NON-SMOOTH, PARTICULARLY COMPOSITE, MULTI-OBJECTIVE OPTIMIZATION

Theory and algorithms

HIROKI TANABE

NON-SMOOTH, PARTICULARLY COMPOSITE, MULTI-OBJECTIVE OPTIMIZATION

Theory and algorithms

by

HIROKI TANABE

Submitted in partial fullfillment of
the requirement for the degree of
DOCTOR OF INFORMATICS
(Applied Mathematics and Physics)

KYOTO UNIVERSITY
KYOTO 606-8501, JAPAN

SEPTEMBER 2022

Preface

To accurately answer all human needs with optimization problems, we cannot avoid considering multi-objective optimization. Humankind is a *greedy* creature that cannot tolerate only a single desire and always has multiple preferences. Unfortunately, many of them conflict with each other, and the best choice to answer all of them seldom exists. This trade-off is what makes multi-objective optimization a tough challenge. Even for problems ideally with multi-objectives, the single-objective models tend to be adopted. However, thanks to the long-standing wisdom of scientists, the development of theories and algorithms for multi-objective optimization has been gradually gaining speed in recent years. As one of the "*dwarfs who ride above the giants*," I would like to contribute to their development, even if only slightly.

This thesis provides theories and algorithms for problems in multi-objective optimization where the objective function is non-smooth. These types of problems are very complex. Thus, it is not practical to consider general non-smooth models for large-scale problems, which have been recently in high demand. Therefore, this thesis mainly focuses on multi-objective optimization problems with a specific structure, called composite models. In detail, this model's every objective function is the sum of differentiable and convex functions. Such models work well, for example, with the loss and regularization models in machine learning.

There are three main contributions of this thesis. One is new merit functions for multi-objective optimization and the elucidation of their properties. A merit function is a function that returns zero in the solution of the problem and a positive number otherwise. We can use it to reformulate the original problem and estimate the rate of convergence of the algorithm. Another contribution is the proximal gradient method for multi-objective optimization problems. It is a first-order method using information from first-order derivatives for composite multi-objective optimization problems. It is more efficient than existing first-order methods for non-smooth multi-objective problems; it has an $O(1/k)$ convergence rate. It can

also generate stationary points for non-convex problems. Another contribution is the accelerated proximal gradient method for multi-objective optimization. It does not work for non-convex problems, but it is faster than the proximal gradient method and solves problems with $O(1/k^2)$. The proposed algorithm is also novel for single-objective problems if the parameters are well-chosen. Its numerical results are better than existing algorithms for single objectives.

Hiroki Tanabe
September 2022

Acknowledgment

This thesis summarizes the author's research during the enrollment in the doctoral course at the Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University. Professor Nobuo Yamashita and Associate Professor Ellen Hidemi Fukuda of the same department, my supervisors, allowed me to conduct this research and provided guidance throughout it. I want to express my deepest gratitude to them.

I likewise thank the members of the System Optimization Laboratory, the scientists of the Operations Research Society of Japan, and many others for their valuable comments and suggestions at the workshops and conferences. I would also like to thank my friends and family for their emotional support.

Part of this research was supported by Grant-in-Aid for JSPS Fellows (20J21961) from the Japan Society for the Promotion of Science.

Contents

List of Figures

List of Tables

List of Symbols and Notations

(x, y) the open line segment between x and y

$[x, y]$ the closed line segment between x and y

$\mathrm{conv}(C)$ the convex hull of C

$\mathrm{dist}\, x, C$ the distance between x and C

$\mathrm{dom}(f)$ the effective domain of function f

$\langle x, y \rangle$ the Euclidean inner product between x and y

$\mathrm{int}(C)$ the interior of C

$\mathcal{J}_f(x)$ the Jacobian matrix of f at x

$\nabla f(x)$ the gradient of f at x

$\|x\|_1$ the ℓ_1-norm of x

$\|x\|_2$ the ℓ_2-norm of x

$\|x\|_\infty$ the ℓ_∞-norm of x

$\partial f(x)$ the subdifferential of f at x

$\mathbf{R}$ the set of real numbers

$\mathbf{R}^n$ the n-dimensional real space

$\mathbf{R}^n_+$ the nonnegative orthant in $\mathbf{R}^n$

Δ^n the unit n-simplex

I_n the $n \times n$ identity matrix

Chapter 1

Introduction

Optimization, a branch of applied mathematics, minimizes (or maximizes) an objective function under given constraints. It is a fundamental technique for operations research and machine learning.

This chapter first describes multi-objective optimization, the subject of this thesis, and composite optimization, a crucial class of non-smooth optimization. It also explains the merit function, an analytical tool for optimization. Finally, it identifies the research challenges on multi-objective optimization problems and explains this thesis's motivations, contributions, and outlines.

1.1 Multi-objective optimization

Optimization problems usually deal with only one objective function. However, many real-world problems have multiple objectives. One solution to this is *multi-objective optimization*, which minimizes several objective functions as follows:

$$\min_{x \in C} \quad F(x), \tag{1.1}$$

where $C \subseteq \mathbf{R}^n$ is a constraint set, and $F \colon \mathbf{R}^n \to (\mathbf{R} \cup \{\infty\})^m$ is a vector-valued function with $F \coloneqq (F_1, \ldots, F_m)^\top$. When $m = 1$, (1.1) reduces to a single-objective optimization. This model has many applications in engineering [**Eschenauer1990**], statistics [**Carrizosa1998**], and machine learning (particularly multi-task learning [**Sener2018**, **Lin2019**] and neural architecture search [**Kim2017**, **Dong2018**, **Elsken2019**]).

In most cases of $m \geq 2$, no single point minimizes all objective functions simul-

taneously, so we use the concept of *Pareto optimality*, a generalization of the usual optimality for single-objective problems. We say that $y \in C$ *Pareto dominates* $x \in C$ if $F_i(y) \leq F_i(x)$ for all $i = 1, \ldots, m$ and $F_j(y) < F_j(x)$ for at least one $j = 1, \ldots, m$, and we call a point *Pareto optimal* if it is not Pareto dominated by any other point. Generally, the Pareto optimal solution is not unique and constitutes a set. We call such a set the *Pareto frontier*. The points in the Pareto frontier are in trade-off relationships with each other, and the decision-makers have to select a solution from it further.

1.1.1 Scalarization approach

The *scalarization approach* [**Gass1955**, **Geoffrion1968**, **Zadeh1963**] is one of the most popular strategies for multi-objective problems. It converts the original multi-objective problem into a parameterized scalar-valued problem.

Let us now introduce the *weighted sum method* [**Zadeh1963**], one of the most well-known scalarization techniques. It scalarize (1.1) with the weight vector $w := (w_1, \ldots, w_m)^\top \in \mathbf{R}^m$ as follows:

$$\min_{x \in \mathbf{R}^n} \quad \langle w, F(x) \rangle, \tag{1.2}$$

where

$$w \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} w_i = 1.$$

When F is convex, for every Pareto optimal solution x^* of (1.1), there exists w such that x^* is the solution of (1.2) [**Miettinen1998**]. However, it may be challenging to choose a *good* weight in advance. Moreover, if F is non-convex, there may be Pareto optimal solutions that are not the solutions of (1.2) for any w, and some w may make (1.2) unbounded.

1.1.2 Heuristics

Heuristics are approaches that do not necessarily lead to the optimal solution but can yield a solution close to the optima at some level. Regarding the multi-objective context, in many cases, heuristics employ evolutionary algorithms, particularly genetic algorithms (GA) such as NSGA-II [**Deb2002**] and NSGA-III [**Deb2014**], being practical for the Pareto frontier enumeration because they are multi-point search

algorithms. These approaches have had some success for real-world problems, but they have the disadvantage that there is no theoretical convergence guarantee to obtain a Pareto solution.

1.1.3 Descent methods

Descent methods [**Fukuda2014**] are iterative algorithms that decrease the objective function values at each iteration. They do not require *a priori* parameters selection like scalarization, and unlike heuristics, we can analyze their global convergence property under reasonable assumptions. All algorithms proposed in this thesis are part of the descent methods. Below we provide typical descent methods for (1.1).

Example 1.1

The steepest descent method [Fliege2000]

Consider a smooth unconstrained multi-objective optimization, i.e., $C = \mathbf{R}^n$ and each F_i is differentiable in (1.1). Then, the steepest descent method updates $\{x^k\}$ by the following operations:

$$\begin{aligned} d^k &:= \underset{d\in\mathbf{R}^n}{\operatorname{argmin}} \left[\max_{i=1,\dots,m} \left\langle \nabla F_i(x^k), d \right\rangle + \frac{1}{2\lambda_k}\|d\|_2^2 \right], \\ x^{k+1} &:= x^k + s_k d^k \end{aligned} \tag{1.3}$$

*with $\lambda_k > 0$ and $s_k > 0$. When $m = 1$, we have $d^k = -\lambda_k \nabla F_1(x)$, which is the steepest descent direction for the scalar optimization [**Cauchy1847**].*

The projected gradient method [Grana-Drummond2004, Fukuda2013]

For a convex-constrained smooth multi-objective optimization, i.e., $C \subseteq \mathbf{R}^n$ is non-empty, closed, and convexand every F_i is differentiable in (1.1), we can use the projected gradient method described by

$$\begin{aligned} z^k &:= \underset{z\in C}{\operatorname{argmin}} \left[\max_{i=1,\dots,m} \left\langle \nabla F_i(x^k), z - x^k \right\rangle + \frac{1}{2\lambda_k}\left\|z - x^k\right\|_2^2 \right], \\ x^{k+1} &:= x^k + s_k \left(z^k - x^k\right) \end{aligned} \tag{1.4}$$

*with $\lambda_k > 0$ and $s_k > 0$. When $m = 1$, (1.4) reduces to the projected gradient method for scalar optimization [**Polyak1963**, **Goldstein1964**, **Goldstein1967**,*

McCormick1969], *i.e.*,

$$z^k := \mathbf{proj}_C(x^k - \lambda_k \nabla F_1(x^k)),$$
$$x^{k+1} := x^k + s_k \left(z^k - x^k\right),$$

where $\mathbf{proj}_C$ *denotes the projection onto* C *given by*

$$\mathbf{proj}_C(x) := \operatorname*{argmin}_{z \in C} \|z - x\|_2. \tag{1.5}$$

Moreover, when $C = \mathbf{R}^n$, (1.4) *amounts to the steepest descent method* (1.3).

The projected subgradient method [Bello-Cruz2013]

Focus on a convex-constrained, non-smooth, and convex multi-objective optimization, i.e., C *is a non-empty, closed, and convex subset of* $\mathbf{R}^n$, *and each* F_i *is convex and non-differentiable in* (1.1). *The subgradient method requires an exogenous sequence* $\{\beta_k\}$ *satisfying*

$$\beta_k > 0, \quad \sum_{k=0}^{\infty} \beta_k = \infty, \quad \textit{and} \quad \sum_{k=0}^{\infty} \beta_k^2 < \infty$$

and generates $\{x^k\}$ *by*

$$x^{k+1} := \operatorname*{argmin}_{z \in C} \left[\frac{1}{2}\|z - x^k\|_2^2 + \frac{\beta_k}{\eta_k} \max_{i=1,\ldots,m} \langle \xi_i^k, z - x^k \rangle\right],$$

where $\xi_i^k \in \partial F_i(x^k)$ *and*

$$\eta_k := \max_{i=1,\ldots,m} \|\xi_i\|.$$

When $m = 1$, *this step represents the projected subgradient method [***Polyak1967**, **Polyak1969**, **Shor1985**, **Alber1998**, **Alber2001**] *for scalar optimization:*

$$x^{k+1} := \mathbf{proj}_C\left(x^k - \frac{\beta_k}{\eta_k}\xi_1^k\right).$$

1.2 Composite optimization

Composite optimization has the following structure:

$$\min_{x \in \mathbf{R}^n} \quad F(x) := f(x) + g(x), \tag{1.6}$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ is L_f-smooth with some $L_f > 0$, and $g\colon \mathbf{R}^n \to \mathbf{R}\cup\{\infty\}$ is closed, proper, and convex. When f is convex, we call (1.6) *convex composite*. This model has many applications, particularly in machine learning. In detail, f and g often represent the loss function and the regularization term, respectively. We list below some typical examples with the structure (1.6).

Example 1.2

Smooth unconstrained minimization

If $g = 0$, (1.6) reduces to the unconstrained smooth minimization

$$\min_{x\in\mathbf{R}^n} \quad f(x),$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ is L_f-smooth.

Convex-constrained smooth minimization

If g is an indicator function of a non-empty, closed, and convex set C, i.e.,

$$g(x) = \delta_C(x) := \begin{cases} 0 & x \in C, \\ \infty & \text{otherwise}, \end{cases} \tag{1.7}$$

then (1.6) amounts to the convex-constrained smooth minimization

$$\min_{x\in C} \quad f(x)$$

with an L-smooth function f.

ℓ_1-regularization

If $g(x) := \tau\|x\|_1$ for some $\tau > 0$, (1.6) reduces to the ℓ_1-regularizaiton

$$\min_{x\in C} \quad f(x) + \tau\|x\|_1$$

with f being L_f-smooth.

1.2.1 The proximal gradient method

The *proximal gradient method* [**Fukushima1981**] is one of the most common algorithms for solving (1.6). For a given $x^0 \in \mathrm{int}(\mathrm{dom}(F))$, it recursively update $\{x^k\}$ by

$$x^{k+1} = \mathbf{prox}_{\lambda g}\left(x^k - \lambda\nabla f(x^k)\right),$$

where **prox** is the *proximal operator* of g with parameter $\lambda > 0$ defined by

$$\mathbf{prox}_{\lambda g}(x) \coloneqq \underset{y \in \mathbf{R}^n}{\operatorname{argmin}} \left[g(x) + \frac{1}{2\lambda} \|x - y\|_2^2 \right].$$

If we can estimate the Lipschitz constant L_f, we can use a constant stepsize $\lambda \in (0, 1/L_f]$. Otherwise, we can determine λ in each iteration by backtracking.

The description of the algorithm now follows.

Algorithm 1.1 The proximal gradient method

Input: $x^0 \in \operatorname{int}(\operatorname{dom}(F)), \varepsilon > 0$

1: $k \leftarrow 0$
2: **repeat**
3: pick $\lambda > 0$
4: $x^{k+1} \leftarrow \mathbf{prox}_{\lambda g}(x^k - \lambda \nabla f(x^k))$
5: $k \leftarrow k + 1$
6: **until** $\|x^k - x^{k-1}\|_\infty < \varepsilon$
7: **return** x^k

With this algorithm, $\{\|x^{k+1} - x^k\|\}$ converges to zero with a rate of $O(\sqrt{1/k})$ and every accumulation point of $\{x^k\}$, if it exists, is stationary point [**Beck2017**]. When f is convex, $\{x^k\}$ converges to the global minima x^*, and $\{F(x^k) - F(x^*)\}$ converges to zero with a rate of $O(1/k)$ [**Beck2017**]. Moreover, when f is strongly convex, $\{x^k\}$ converges linearly to x^* [**Beck2017**]. Furthermore, if we assume the so-called proximal-PL condition, $\{F(x^k)\}$ converges linearly to $F(x^*)$ [**Karimi2016**].

1.2.2 The accelerated proximal gradient method

When f is convex, the accelerated proximal gradient method, also known as Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [**Beck2009**], can solve (1.6) with an $O(1/k^2)$ rate of convergence, while the proximal gradient method achieves a rate of $O(1/k)$.

We describe below the algorithm. Like the proximal gradient method, the stepsize λ may be fixed at a constant or updated by the backtracking procedure.

With Algorithm 1.2, $\{F(x^k) - F(x^*)\}$ for the global minima x^* converges to zero with a rate of $O(1/k^2)$ [**Beck2009**], but the convergence of iterates remains unknown. With a slite modification, that is, changing the update rule of the momentum factor with $t_k = (k + a - 1)/a$ for some $a > 2$, we can prove that $\{x^k\}$ converges to the minima while keeping the convergence rate of $O(1/k^2)$.

Algorithm 1.2 The accelerated proximal gradient method

Input: $x^0 \in \text{int}(\text{dom}(F)), \varepsilon > 0$
$k \leftarrow 1$
$y^1 \leftarrow x^0$
$t_1 \leftarrow 1$
repeat
    pick $\lambda > 0$
    $x^k \leftarrow \mathbf{prox}_{\lambda g}(y^k - \lambda \nabla f(y^k)$
    $t_{k+1} \leftarrow \sqrt{t_k^2 + 1/4} + 1/2$
    $\gamma_k \leftarrow (t_k - 1)/t_{k+1}$
    $y^{k+1} \leftarrow x^k + \gamma_k(x^k - x^{k-1})$
    $k \leftarrow k + 1$
until $\|x^k - y^k\|_\infty < \varepsilon$
return x^k

1.3 Merit functions

Merit functions [**Fukushima1996**] are maps that return zeros at the problems' solutions and strictly positive values otherwise. In other words, they are the objective functions of optimization problems that have the same solutions as the original problems. Therefore, it is desirable for the merit functions to have the following properties:

- Quick computability;
- Continuity;
- Differentiability;
- Optimality of the stationary points;
- Level-boundedness;
- Error-boundedness.

Moreover, as we can consider the merit functions to represent how far feasible points are from the optimal solutions, they are useful for analyzing convergence rates of the optimization algorithm.

1.3.1 Merit functions for variational inequalities

Merit functions have evolved in the context of reformulating variational inequalities (VIs) and complementarity problems (CPs) as optimization problems [**Fukushima1996**]. The *variational inequality* (VI) consists in finding $x \in C$ such that

$$\langle T(x), y - x\rangle \geq 0 \quad \text{for all } y \in C, \tag{1.8}$$

where $C \subseteq \mathbf{R}^n$ is nonempty, closed, and convex, and $T\colon \mathbf{R}^n \to \mathbf{R}^n$ is continuous. We can also rewrite (1.8) as the following *complementarity problem* (CP):

$$T(x) \geq 0, \quad x \geq 0, \quad \text{and} \quad \langle T(x), x\rangle \geq 0. \tag{1.9}$$

In particular, if T is affine, the we call (1.9) the *linear complementarity problem* (LCP). There are many merit functions for VIs and CPs, but here we illustrate the most basic two merit functions for VIs.

Example 1.3 (Merit functions for the VI (1.8))

The classical gap function [Auslender1976, Hearn1982]

We call the function $G_0\colon \mathbf{R}^m \to \mathbf{R} \cup \{\infty\}$ *the* classical gap function*:*

$$G_0(x) \coloneqq \sup_{y \in C} \langle T(x), x - y\rangle. \tag{1.10}$$

It has the following properties:

- $G_0(x) \geq 0$ *for all* $x \in C$*;*
- $G_0(x) = 0$ *and* $x \in C$ *if and only if* x *satisfies* (1.8)*;*
- *If* C *is bounded,* G_0 *is finite everywhere.*

The top two indecate that G_0 *is a merit function for the VI* (1.8)*.*

The regularized gap function [Fukushima1992, Auchmuty1989]

For a given parameter $\alpha > 0$*, we can consider the* regularized gap function $G_\alpha\colon \mathbf{R}^n \to \mathbf{R}$ *defined by*

$$G_\alpha(x) \coloneqq \max_{y \in C} \left[\langle T(x), x - y\rangle - \frac{\alpha}{2}\|x - y\|_2^2\right], \tag{1.11}$$

which is a merit function for the VI (1.8)*, too. Since* (1.11) *maximizes a strongly concave function on a nonempty, closed, and convex set, even if* C

is unbounded, an unique point attains the maximum, and G_α is finite everywhere. Moreover, denoting such a maximizer by $H_\alpha(x)$, if T is continuously differentiable, G_α is also differentiable at any point x, and we have

$$\nabla G_\alpha(x) = T(x) - [\mathcal{J}_T(x) - \alpha I_n](H_\alpha(x) - x).$$

Note that

$$H_\alpha(x) = \mathbf{proj}_C(x - \alpha^{-1}T(x)).$$

Furthermore, if the Jacobian $\mathcal{J}_T(x)$ is positive definite on C, any stationary point of the problem

$$\min_{x \in C} \quad G_\alpha(x)$$

*solves the VI (1.8) [**Fukushima1992**]. In addition, if T is strongly monotone with modulus $\mu > 0$, i.e.,*

$$\langle T(x) - T(x'), x - x' \rangle > \mu \|x - x'\|^2 \quad \text{for all } x, x' \in \mathbf{R}^n,$$

*and if $\alpha < 2\mu$, then G_α has the following error bound property [**Taji1993**]:*

$$\|x - x^*\| \leq \sqrt{\frac{G_\alpha(x)}{\mu - \alpha/2}} \quad \text{for all } x \in S,$$

where x^* is the unique solution of the VI (1.8).

1.3.2 Merit functions for multi-objective problems

The history of research on merit functions for multi-objective problems are relatively new, beginning in 1998 with **Chen1998** on (1.1) under the assumptions of polyhedrality of C and convexity of F. Afterward, various merit functions have appeared for multi-objective problems, including multi-objective optimization [**Liu2009**, **Dutta2017**], (finite-dimensional) vector variational inequalities [**Chen2000**, **Konnov2005**, **Li2005**, **Yang2002**, **Yang2003**, **Charitha2010**, **Li2010**], and (finite-dimensional) vector equilibrium problems [**Huang2007**, **Li2005**, **Li2007**, **Li2006**, **Mastroeni2003**]. Below we pick up generalizations of Example 1.3 to the *weak Stamnpacchia type vector variational inequality* $(SVVI)^w$, which consists in

finding $x \in C$ such that

$$(\langle T_1(x), y - x\rangle, \ldots, \langle T_m(x), y - x\rangle) \notin -\operatorname{int}(\mathbf{R}^m_+) \quad \text{for all } y \in C, \tag{1.12}$$

where $C \subseteq \mathbf{R}^n$ is a nonempty, closed, convex, and $T_i \colon \mathbf{R}^n \to \mathbf{R}^n, i = 1, \ldots, m$. Note that x satisfies (1.12) if and only if x is weakly Pareto optimal for (1.1) when F_i is differentiable and $T_i = \nabla F_i$ for each $i = 1, \ldots, m$.

Example 1.4
The gap function for $(SVVI)^w$ **[Charitha2010, Li2010]**

We can write the gap function $G_0 \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ *for* (1.12) *as*

$$G_0(x) \coloneqq \min_{\lambda \in \Delta^m} \sup_{y \in C} \left\langle \sum_{i=1}^m \lambda_i T_i(x), x - y \right\rangle.$$

When $m = 1$, *it correpsponds to* (1.10). *Like* (1.10), G_0 *is a merit function for* (1.12), *i.e.,*

- $G_0(x) \geq 0$ *for all* $x \in C$;
- $G_0(x) = 0, x \in C$ *if and only if* x *solves* (1.12),

and it is finite-valued if C *is bounded.*

The regularized gap function for $(SVVI)^w$ **[Charitha2010]**

We can define the regularized gap function $G_\alpha \colon \mathbf{R}^n \to \mathbf{R}$ *with* $\alpha > 0$ *for* (1.12) *by*

$$G_\alpha(x) \coloneqq \min_{\lambda \in \Delta^m} \max_{y \in C} \left[\left\langle \sum_{i=1}^m \lambda_i T_i(x), x - y \right\rangle - \frac{\alpha}{2} \|x - y\|_2^2 \right],$$

matching (1.11) *when* $m = 1$. *It also satisfies the two propeties as a merit function for* (1.12). *Moreover, if each* $T_i, i = 1, \ldots, m$ *is continuously differentiable, then* G_α *is directionally differentiable in any direction* $d \in \mathbf{R}^n$, *and*

$$G'_\alpha(x; d) = \min_{\lambda \in \Lambda(x)} \left[\left\langle \sum_{i=1}^m \lambda_i T_i(x) - \sum_{i=1}^m \lambda_i \mathcal{J}_{T_i}(x)(H_\alpha(x, \lambda) - x), d \right\rangle + \alpha \langle H_\alpha(x, \lambda) - x, d \rangle \right],$$

where

$$H_\alpha(x,\lambda) \coloneqq \mathbf{proj}_C\left(x - \alpha^{-1}\sum_{i=1}^m \lambda_i T_i(x)\right),$$
$$T_\alpha(x,\lambda) \coloneqq -\left\langle \sum_{i=1}^m \lambda_i T_i(x), H_\alpha(x,\lambda) - x\right\rangle - \frac{\alpha}{2}\|H_\alpha(x,\lambda)-x\|_2^2,$$
$$\Lambda(x) \coloneqq \{\lambda \in \Delta^m \mid G_\alpha(x) = T_\alpha(x,\lambda)\}.$$

Particularly, if $\Lambda(x)$ is a singleton, i.e., $\Lambda(x) = \{\lambda(x)\}$, G_α is Gateaux differentiable at x and

$$\nabla G_\alpha(x)$$
$$= \sum_{i=1}^m \lambda(x)_i T_i(x) - \sum_{i=1}^m \lambda(x)_i \mathcal{J}_{T_i}(x)[H_\alpha(x,\lambda(x)) - x] + \alpha[H_\alpha(x,\lambda(x)) - x].$$

Furthermore, if each $T_i, i = 1,\dots,m$ is strongly monotone with modulus $\mu_i > 0$, and if $\alpha < 2\mu$ with $\mu \coloneqq \min_{i=1,\dots,m}\mu_i$, then G_α provides the error bound:

$$\mathrm{dist}(x, sol(SVVI)^w) \le \sqrt{\frac{G_\alpha(x)}{\mu - \alpha/2}} \quad \textit{for all } x \in C,$$

where $sol(SVVI)^w$ denotes the solution set of (1.12).

1.4 Motivations and contributions

As discussed in Section 1.1, multi-objective optimization (1.1) is an indispensable model in dealing with real-world problems, and the studies on its theories and algorithms have great significance. On the other hand, many previous studies on multi-objective optimization, particularly on the descent methods described in Section 1.1.3 and the merit functions described in Section 1.3.2, have dealt with smooth problems, and there is still room for exploration for non-smooth problems. The projected subgradient method inroduced in Example 1.1 can handle non-smooth multi-objective optimization, but it may not work well for large-scale problems due to the stepsize decay.

This thesis focuses on non-smooth multi-objective optimization problems with

specific structures, mainly the generalization of the composite model introduced in Section 1.2, i.e., (1.1) with

$$F_i(x) = f_i(x) + g_i(x) \quad \text{for all } i = 1, \dots, m, \tag{1.13}$$

where f_i is continuously differentiable and g_i is closed, proper, and convex. Then, we presents their theory and algorithms.

1.5 Outline of the thesis

After introducing in Chapter 2 some symbols, basic definitions, and their properties necessary for the discussion, Chapter 3 proposes and characterizes three new types of merit functions for non-smooth multi-objective optimization problems: the gap function for continuous problems, the regularized gap function for convex problems, and the regularized and partially linearized gap functions for composite problems. Chapter 4 develops the proximal gradient method for composite multi-objective optimization problems, describes its convergence, convergence rate, applications to robust multi-objective optimization, and performs numerical experiments. Chapter 5 presents its acceleration applicable with *convex* composite objectives: the accelerated proximal gradient method or Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) and provides similar discussions. We note here that our multi-objective FISTA represents a new algorithm even for single objectives, depending on the choice of acceleration factors, and performs better in numerical experiments.

Chapter 2

Preliminaries

This chapter presents some notations, basic definitions, and their properties used in this thesis.

2.1 Vectors and matrices

Let $\mathbf{R}^p$ denote the space of p-dimensional real column vectors. Meanwhile, we write the set of real numbers as simply $\mathbf{R}$ or $(-\infty, +\infty)$ instead of $\mathbf{R}^1$. Moreover, $\mathbf{R}^{q\times p}$ stand for the space formed by $q\times p$ real matrices. In addition, define the non-negative orthant in $\mathbf{R}^p$ by

$$\mathbf{R}^p_+ \coloneqq \{v \in \mathbf{R}^p \mid v_i \geq 0, i = 1, \ldots, p\},$$

and define the unit simplex $\Delta^p \subseteq \mathbf{R}^p_+$ by

$$\Delta^p \coloneqq \left\{ v \in \mathbf{R}^p_+ \;\middle|\; \sum_{i=1}^p v_i = 1 \right\}. \tag{2.1}$$

The orthant $\mathbf{R}^p_+$ induces the partial orders for any $v^1, v^2 \in \mathbf{R}^p$: $v^1 \leq v^2$ (alternatively, $v^2 \geq v^1$ if $v^2 - v^1 \in \mathbf{R}^p_+$, and $v^1 < v^2$ (alternatively, $v^2 > v^1$) if $v^2 - v^1 \in \mathrm{int}(\mathbf{R}^p_+)$. In other words, we say that $v^1 \leq (<)\ v^2$ if $v^1_i \leq (<)\ v^2_i$ for all $i = 1, \ldots, p$. Furthermore, let $\langle \cdot, \cdot \rangle$ stand for the Euclidean inner product, i.e., $\langle v^1, v^2 \rangle \coloneqq \sum_{i=1}^p v^1_i v^2_i$. We also define ℓ_2-norm $\|\cdot\|_2$, ℓ_1-norm $\|\cdot\|_1$, and ℓ_∞-norm $\|\cdot\|_\infty$ by

$$\|v\|_2 \coloneqq \sqrt{\langle v, v \rangle} = \sum_{i=1}^p v_i^2, \quad \|v\|_1 \coloneqq \sum_{i=1}^p |v_i|, \quad \text{and} \quad \|v\|_\infty \coloneqq \max_{i=1,\ldots,p} |v_i|$$

for any $v \in \mathbf{R}^p$.

2.2 Convexity and semi-continuity

We first define the convexity of sets and functions. A set $C \subseteq \mathbf{R}^p$ is *convex* if

$$(1-\alpha)v^1 + \alpha v^2 \in C \quad \text{for all } v^1, v^2 \in C, \alpha \in [0,1].$$

Likewise, a function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is *convex* if

$$h((1-\alpha)x + \alpha y) \leq (1-\alpha)h(x) + \alpha h(y) \quad \text{for all } x, y \in \mathrm{dom}(h), \alpha \in [0,1],$$

strictly convex if

$$h((1-\alpha)x + \alpha y) < (1-\alpha)h(x) + \alpha h(y) \quad \text{for all } x, y \in \mathrm{dom}(h), \alpha \in (0,1),$$

and μ_f*-convex* with $\mu_f \in \mathbf{R}$ if

$$h((1-\alpha)x + \alpha y) \leq (1-\alpha)h(x) + \alpha h(y) - \frac{\mu_f}{2}\alpha(1-\alpha)\|x-y\|^2 \quad \text{for all } x, y \in \mathrm{dom}(h), \alpha \in [0,1],$$

where $\mathrm{dom}(h)$ stands for the *effective domain* of h given by

$$\mathrm{dom}(h) \coloneqq \{x \in \mathbf{R}^p \mid h(x) < \infty\}.$$

In particular, the *strong convexity* (with modulus μ_f) denotes the μ_f-convexity with $\mu_f > 0$. We also note that the 0-convexity is equivalent to the usual convexity. Moreover, if $\mathrm{dom}(h) \neq \emptyset$ for some convex function $h\colon (-\infty, +\infty]$, we say that h is *proper* and convex. On the other hand, we call h to be concave if $-h$ is convex. Every definition and argument relating convex functions also holds for concave functions by appropriately interchenging $\leq$ and $\geq$, $+\infty$ and $-\infty$, sup and inf, etc.

Let us now introduce the semi-continuity of functions. For all $\{x^k\} \subseteq \mathbf{R}^p$ converging to $x \in \mathbf{R}^p$, a function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is *upper semi-continuous* if

$$h(x) \geq \limsup_{k \to \infty} h(x^k)$$

and *lower semi-continuous* if

$$h(x) \leq \liminf_{k\to\infty} h(x^k).$$

A necessary and sufficient condition for h to be lower semi-continuous is that the level set $\mathbf{lev}_\alpha(h)$ given by

$$\mathbf{lev}_\alpha(h) \coloneqq \{x \in \mathbf{R}^p \mid h(x) \leq \alpha\} \tag{2.2}$$

is closed for any $\alpha \in \mathbf{R}$. We refer to lower-semicontinuous, proper, and convex functions as *closed, proper, and convex* functions. The level sets of convex functions are convex, and the level sets of closed, proper, and convex functions are closed and convex.

2.3 Differentiability

Suppose that $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is finite-valued in an appropriate neighborhood of $x \in \mathbf{R}^p$. If h has the partial derivative

$$\frac{\partial h(x)}{\partial x_i} \coloneqq \lim_{\alpha\to 0} \frac{h(x+\alpha e^i) - h(x)}{\alpha} \quad \text{for all } i = 1, \ldots, p$$

with e^i being the unit vector along the x_i-axis, and if

$$h(x+\varepsilon) = h(x) + \langle \nabla h(x), \varepsilon \rangle + o(\|\varepsilon\|_2) \quad \text{for all } \varepsilon \in \mathbf{R}^p$$

with $o\colon [0, +\infty) \to \mathbf{R}$ satisfying $\lim_{\alpha\to 0} o(\alpha)/\alpha$ and

$$\nabla h(x) \coloneqq \begin{bmatrix} \dfrac{\partial h(x)}{\partial x_1} \\ \vdots \\ \dfrac{\partial h(x)}{\partial x_p} \end{bmatrix},$$

then h is *differentiable* at x, and we call $\nabla h(x) \in \mathbf{R}^p$ a *gradient* of h at x. Particularly, if $\nabla h(x)$ is continuous at x, we say that h is *continuously differentiable* at x.

Again, if h has second-order derivatives and

$$h(x+\varepsilon) = h(x) + \langle \nabla h(x), \varepsilon \rangle + \frac{1}{2}\langle \varepsilon, \nabla^2 h(x)\varepsilon \rangle + o(\|h\|_2^2)$$

with

$$\nabla^2 h(x) \coloneqq \begin{bmatrix} \dfrac{\partial^2 h(x)}{\partial x_1\, \partial x_1} & \cdots & \dfrac{\partial^2 h(x)}{\partial x_1\, \partial x_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 h(x)}{\partial x_p\, \partial x_1} & \cdots & \dfrac{\partial^2 h(x)}{\partial x_p\, \partial x_p} \end{bmatrix},$$

then h is *twice differentiable* at $x \in \mathbf{R}^p$, and $\nabla^2 h(x)$ is a *Hessian matrix* of h at x. When $\nabla^2 h$ is continuous at x, h is *twice continuously differentiable* at x, and then $\nabla^2 h(x)$ is symmetric. On the other hand, for a vector-valued function $h\colon \mathbf{R}^p \to \mathbf{R}^q$ with $h \coloneqq (h_1, \ldots, h_m)^\top$, $\mathcal{J}_h(x)$ denotes the Jacobian matrix of h at x, that is,

$$\mathcal{J}_h(x) \coloneqq \begin{bmatrix} \dfrac{\partial h_1(x)}{\partial x_1} & \cdots & \dfrac{\partial h_1(x)}{\partial x_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_q(x)}{\partial x_1} & \cdots & \dfrac{\partial h_q(x)}{\partial x_p} \end{bmatrix} = [\nabla h_1(x), \ldots, \nabla h_q(x)]^\top \in \mathbf{R}^{q\times p}, \tag{2.3}$$

where $\top$ denotes transpose.

2.4 Hölder and Lipschitz continuity

We call $h\colon \mathbf{R}^p \to \mathbf{R}$ to be *locally Hölder continuous* with exponent $\alpha > 0$ if for every bounded set $\Omega \subseteq \mathbf{R}^p$ there exists $L_h > 0$ such that

$$|h(x) - h(y)| \le L_h \|x - y\|_2^\alpha \quad \text{for all } x, y \in \Omega.$$

In particular, when L_h does not depend on Ω, we say that h is Hölder continuous with exponent $\alpha > 0$. Moreover, we refer to the (local) Hölder continuity with exponent 1 as the *(local) Lipschitz continuity*. When h is Lipschitz continuous, we call L_h the *Lipschitz constant*, and we also say that h is L_h*-Lipschitz continuous*. As the following lemma shows, many functions with *good* properties are locally Lipschitz continuous.

Lemma 2.1
Continuously differentiable functions and finite-valued convex functions are locally Lipschitz continuous.

Proof. The former is due to the mean value theorem, and the latter is from [**WayneStateUniversity1**

Furthermore, if h is continuously differentiable and ∇h is L_h-Lipschitz continuous, we say that h is L_h-smooth. We now recall the so-called descent lemma [**Bertsekas1999**] as follows:

Lemma 2.2 (Descent Lemma [Bertsekas1999])
Let $h\colon \mathbf{R}^p \to \mathbf{R}$ is L_h-smooth on $\mathbf{R}^p$ with $L_h > 0$. Then, we have

$$|h(y) - h(x) - \langle \nabla h(x), y - x\rangle| \leq \frac{L_h}{2}\|x - y\|_2^2 \quad \textit{for all } x, y \in \mathbf{R}^p.$$

2.5 Directional derivatives and subgradients

A function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is *directionally differentiable* at $x \in \mathbf{R}^p$ in a direction $d \in \mathbf{R}^p$ if

$$h'(x; d) \coloneqq \lim_{\alpha \searrow 0} \frac{h(x + \alpha d) - h(x)}{\alpha} \tag{2.4}$$

exists, and then we call $h'(x; d)$ the *directional derivative* at x in a direction d. When h is differentiable at x, we have $h'(x; d) = \langle \nabla h(x), d\rangle$ for all $d \in \mathbf{R}^p$. As the following lemma implies, convex functions are directionally differentiable if we allow $\pm\infty$ as a limit.

Lemma 2.3 ([Bertsekas2003])
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be convex. Then, the function $h_{x,d}\colon (0, +\infty) \to (-\infty, +\infty]$ defined by

$$h_{x,d}(\alpha) \coloneqq \frac{h(x + \alpha d) - h(x)}{\alpha}$$

is non-decreasing. In particular, it follows that

$$h'(x; d) \leq h_{x,d}(\alpha) \leq h(x + d) - h(x) \quad \textit{for all } x, d \in \mathbf{R}^p, \alpha \in (0, 1].$$

On the other hand, for a proper and convex function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$, we call $\xi \in \mathbf{R}^p$ a *subgradient* of h at $x \in \mathbf{R}^p$ if

$$h(y) - h(x) \geq \langle \xi, y - x\rangle \quad \text{for all } y \in \mathbf{R}^p, \tag{2.5}$$

and we write $\partial h(x)$ the *subdifferential* of h at x, i.e., the set of all subgradients of h at x. When h is differentiable at x, $\partial h(x)$ amounts to a singular $\{\nabla h(x)\}$.

2.6 The proximal operator and Moreau envelope

We suppose that $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is closed, proper, and convex. Then, we define the *Moreau envelope* or *Moreau-Yosida regularization* $\mathcal{M}_h\colon \mathbf{R}^p \to \mathbf{R}$ by

$$\mathcal{M}_h(x) \coloneqq \min_{y \in \mathbf{R}^p} \left[h(y) + \frac{1}{2}\|x - y\|_2^2 \right]. \tag{2.6}$$

The minimization problem in (2.6) has a unique solution because of the strong convexity of its objective function. We call this solution the *proximal operator* and write it as

$$\mathbf{prox}_h(x) = \operatorname*{argmin}_{y \in \mathbf{R}^p} \left[h(y) + \frac{1}{2}\|x - y\|_2^2 \right]. \tag{2.7}$$

The proximal operator is non-expansive, i.e., $\|\mathbf{prox}_h(x) - \mathbf{prox}_h(y)\|_2 \leq \|x - y\|_2$ for any $x, y \in \mathbf{R}^p$. This also means that $\mathbf{prox}_h$ is 1-Lipschitz continuous. Moreover, when h is the indicator function (1.7) of a non-empty, closed, and convex set $C \subseteq \mathbf{R}^p$, we have

$$\mathbf{prox}_{\delta_C}(x) = \mathbf{proj}_C(x), \tag{2.8}$$

where $\mathbf{proj}_C$ is the projection onto C defined by (1.5). Even if h is non-differentiable, its Moreau envelope $\mathcal{M}_h$ is differentiable.

Theorem 2.4 ([Beck2017])
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be closed, proper, and convex. Then, $\mathcal{M}_h$ is 1-smooth and

$$\nabla \mathcal{M}_h(x) = x - \mathbf{prox}_h(x).$$

We also refer to the so-called second prox theorem as well as a corollary quickly derived from it.

Theorem 2.5 (Second prox theorem [Beck2017])
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be closed, proper, and convex. Then, it follows that

$$\langle x - \mathbf{prox}_h(x), y - \mathbf{prox}_h(x) \rangle \leq h(y) - h(\mathbf{prox}_h(x)) \quad \textit{for all } x, y \in \mathbf{R}^p.$$

Corollary 2.6
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be closed, proper, and convex. Then, we have

$$\|x - \mathbf{prox}_h(x)\|_2^2 \leq h(x) - h(\mathbf{prox}_h(x)) \quad \textit{for all } x \in \mathbf{R}^p.$$

2.7 Stability and sensitivity analysis

We consider the following parameterized optimization problem:

$$\min_{x \in X} \quad h(x, \xi), \tag{2.9}$$

depending on the parameter vector $\xi \in \Xi$. We assume that $X \subseteq \mathbf{R}^p$ and $\Xi \subseteq \mathbf{R}^q$ are non-empty and closed. Let us write the optimal value function of (2.9)

$$\phi(\xi) \coloneqq \inf_{x \in X} h(x, \xi) \tag{2.10}$$

and the associated set as

$$\Phi(\xi) \coloneqq \{x \in X \mid \phi(\xi) = h(x, \xi)\}.$$

The following proposition describes the directional differentiability of the optimal value function ϕ.

Proposition 2.7 ([Bonnans2000])
Let $\xi^0 \in \Xi$. Suppose that

(i) the function $h(x, \xi)$ is continuous on $X \times \Xi$;

(ii) there exist $\alpha \in \mathbf{R}$ and a compact set $C \subseteq X$ such that for every $\hat{\xi}$ near ξ^0, the level set $\mathbf{lev}_\alpha\, h(\cdot, \hat{\xi})$ is non-empty and contained in C;

(iii) for any $x \in X$, the function $h_x(\cdot) \coloneqq h(x, \cdot)$ is directionally differentiable at ξ^0;

(iv) if $\xi \in \Xi, t_k \searrow 0$, and $\{x^k\} \subseteq C$, then $\{x^k\}$ has an accumulation point $\bar{x}$ such that

$$\limsup_{k \to \infty} \frac{h(x^k, \xi^0 + t_k(\xi - \xi^0)) - h(x^k, \xi^0)}{t_k} \geq h'_{\bar{x}}(\xi^0; \xi - \xi^0).$$

Then, the optimal value function ϕ given by (2.10) is directionally differentiable at ξ^0

and

$$\phi'(\xi^0;\xi-\xi^0)=\inf_{x\in\Phi(\xi^0)}h'_x(\xi^0;\xi-\xi^0).$$

2.8 Pareto optimality

Let us introduce the concept of optimality for the multi-objective optimization problem (1.1).

Definition 2.1 (Pareto optimality and weak Pareto optimality)
For (1.1), we say that $x\in C$ is

(i) Pareto optimal *if there is no $y\in C$ such that $F(y)\leq F(x)$ and $F(y)\neq F(x)$;*

(ii) weakly Pareto optimal *if there does not exist $y\in C$ such that $F(y)<F(x)$.*

By definition, weak Pareto optimality containss Pareto optimality, though both definitions reduce to the usual optimality when $m=1$. On the other hand, if the objective functions are non-convex, it is challenging to find Pareto minima or weak Pareto minima. In such cases, optimization algorithms basically aim to get Pareto stationary points defined as follows:

Definition 2.2 (Pareto stationarity)
Assume that F_i is directionally differentiable for every $i=1,\dots,m$ and C is nonempty, closed, and convex. Then, we call $x\in C$ Pareto stationary *if*

$$\max_{i=1,\dots,m}F'_i(x;y-x)\geq 0\quad \textit{for all } y\in C.$$

We state below the relation among the three concepts given by Definitions 2.1 and 2.2

Lemma 2.8
Suppose that F_i is directionally differentiable for every $i=1,\dots,m$ and C is nonempty, closed, and convex. Then, the following three claims hold.

(i) If $x\in C$ is weakly Pareto optimal for (1.1), then x is Pareto stationary for (1.1).

(ii) Let every $F_i, i=1,\dots,m$ be convex. Then, all Pareto stationary points of (1.1) are weakly Pareto optimal for (1.1).

(iii) Suppose that F_i is strictly convex for any $i = 1, \dots, m$. Then, every Pareto stationary points of (1.1) is Pareto optimal for (1.1).

Proof. We prove each claim's contraposition.

Claim (i) : Assume that $x \in C$ is not Pareto stationary. Then, Definition 2.2 shows that for some $y \in C$ we have $\max_{i=1,\dots,m} F_i'(x; y - x) < 0$. By the definition (2.4) of the directional derivative, for a sufficiently small scalar $\alpha > 0$, we obtain

$$\max_{i=1,\dots,m} [F_i(x + \alpha(y - x)) - F_i(x)] < 0,$$

which means that x is not weakly Pareto optimal from Definition 2.1 (ii).

Claim (ii) : Suppose that $x \in C$ is not weakly Pareto optimal. Then, Definition 2.1 (ii) implies that there exists $y \in C$ such that $F_i(y) < F_i(x)$ for all $i = 1, \dots, m$. Therefore, the convexity of F_i and Lemma 2.3 give

$$F'(x; y - x) \leq F_i(y) - F_i(x) < 0 \quad \text{for all } i = 1, \dots, m.$$

Hence, we get

$$\max_{i=1,\dots,m} F'(x; y - x) < 0,$$

which implies that x is not Pareto stationary from Definition 2.2.

Claim (iii) : Suppose that $x \in C$ is not Pareto optimal. From Definition 2.1 (i), there exists $y \in C$ such that $F(y) \leq F(x)$ and $F(y) \neq F(x)$. Since F_i is strictly convex for every $i = 1, \dots, m$, we have

$$F(x + \alpha(y - x)) < F(x) + \alpha(F(y) - F(x)) \quad \text{for all } \alpha \in (0, 1).$$

Reducing $F(x)$ and dividing by α from both sides lead to

$$\frac{F(x + \alpha(y - x)) - F(x)}{\alpha} < F(y) - F(x) \leq 0.$$

Applying Lemma 2.3 to each component yields

$$F_i'(x; y - x) < F_i(y) - F_i(x) \leq 0,$$

which shows that x is not Pareto stationary. □

Chapter 3

Merit functions for multi-objective optimization

3.1 Introduction

This chapter considers the convex-constrained multi-objective optimization problems, i.e., (1.1) with C being non-empty, closed, and convex. It presents new merit functions for them, and discusses their properties mentioned in Section 1.3.

In detail, it proposes the following three merit functions for (1.1):

(i) the gap function for continuous multi-objective optimization;

(ii) the regularized gap function for convex multi-objective optimization;

(iii) the regularized and partially linearized gap function for composite multi-objective optimization.

In Table 3.1, we summarize the properties of those merit functions, which will be shown in the subsequent sections. There, 'Sol.' represents the types of Pareto solutions for (1.1) corresponding to the minima (zero points) of the merit functions. Moreover, 'SP,' 'LB,' and 'EB' indicate each F_i's sufficient conditions so that stationary points of the merit functions can solve (1.1), the merit functions are level-bounded, and the merit functions provide error bounds, respectively. The gap function (i) connects its minima and the weak Pareto solutions of (1.1) but does not have good properties in other aspects. The regularized gap function (ii) has better properties but requires the convexity of F_i. The regularized and partially lin-

earized gap function (iii) relaxes the convexity assumption and is easy to compute for particular problems.

Table 3.1: Properties of our proposed merit functions

(a) Proposed merit functions and their properties

	Obj.	Sol.	Cont.	Diff.	SP	LB	EB
(i)	Cont.	WPO	LSC	×	×	LB	SgC
(ii)	Conv.		Cont.	DD	SC	Conv., LB	
(iii)	Comp.	PS			SC, C^2	Conv., LB, etc.	SgC, etc.

(b) Table of abbreviations

Obj.	Objective functions
Sol.	Solutions
Cont.	Continuity
Diff.	Differentiability
SP	Statioary points
LB	Level-boundedness
EB	Error bounds
Cont.	Continuity
Comp.	Composite
WPO	Weak Pareto optimality
PS	Pareto stationarity
LSC	Lower semicontinuity
DD	Directional differentiability
C^2	Twice continuously differentiable
SC	Strict convexity
SgC	Strong convexity

We summarize the structure of the rest of this chapter. Section 3.2 proposes different merit functions for multi-objective optimization with continuous objectives, convex objectives, and composite objectives, along with methods for evaluating the function values, the differentiability, and the stationary point properties. Furthermore, sufficient conditions for them to be level-bounded and to provide error bounds are given in Sections 3.4 and 3.5, respectively.

3.2 Merit functions and their basic properties

This section proposes different types of merit functions for the multi-objective optimization (1.1), considering three cases: when the objective function F is continuous, when it is convex, and when it has a composite structure.

3.2.1 A gap function for continuous multi-objective optimization

First, we assume only continuity on F other than continuity and propose a gap function $u_0 \colon C \to (-\infty, +\infty]$ as follows:

$$u_0(x) := \sup_{y \in C} \min_{i=1,\dots,m} [F_i(x) - F_i(y)]. \tag{3.1}$$

When F is linear, this merit function has already been discussed in [**Liu2009**], but here we consider the more general nonlinear cases. We now show that u_0 is a merit function in the sense of weak Pareto optimality.

Theorem 3.1
Let u_0 be defined by (3.1). Then, we have $u_0(x) \geq 0$ for all $x \in C$. Moreover, $x \in C$ is weakly Pareto optimal for (1.1) if and only if $u_0(x) = 0$.

Proof. Let $x \in C$. By the definition (3.1) of u_0, we get

$$u_0(x) = \sup_{y \in C} \min_{i=1,\dots,m} [F_i(x) - F_i(y)] \geq \min_{i=1,\dots,m} [F_i(x) - F_i(x)] = 0.$$

On the other hand, again considering the definition (3.1) of u_0, we obtain

$$u_0(x) = 0 \iff \min_{i=1,\dots,m} [F_i(x) - F_i(y)] \leq 0 \quad \text{for all } y \in C.$$

This is equivalent to the existence of $i = 1, \dots, m$ such that

$$F_i(x) - F_i(y) \leq 0 \quad \text{for all } y \in C,$$

i.e., there does not exist $y \in C$ such that

$$F_i(x) - F_i(y) > 0 \quad \text{for all } i = 1, \dots, m,$$

which means that x is weakly Pareto optimal for (1.1) by Definition 2.1 (i). □

The following theorem is clear from the continuity of F_i.

Theorem 3.2
The function u_0 defined by (3.1) is lower semicontinuous on C.

Theorems 3.1 and 3.2 imply that if $u_0(x^k) \to 0$ holds for some bounded sequence $\{x^k\}$, its accumulation points are weakly Pareto optimal. Thus, we can use u_0 to measure the complexity of multi-objective optimization methods.

Moreover, Theorem 3.1 implies that we can get weakly Pareto optimal solutions via the following single-objective optimization problem:

$$\min_{x \in C} \quad u_0(x).$$

However, if F_i is not bounded from below on S, we cannot guarantee that u_0 is finite-valued. Moreover, even if u_0 is finite-valued, u_0 does not preserve the differentiability of the original objective function F.

3.2.2 A regularized gap function for convex multi-objective optimization

Here, we suppose that each component F_i of the objective function F of (1.1) is convex. Then, we define a regularized gap function $u_\alpha \colon C \to \mathbf{R}$ with a given constant $\alpha > 0$, which overcomes the shortcomings mentioned at the end of the previous subsection, as follows:

$$u_\alpha(x) := \max_{y \in C} \min_{i=1,\dots,m} \left[F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \right]. \tag{3.2}$$

Note that the strong concavity of the function inside $\max_{y \in C}$ implies that u_α is finite-valued and there exists a unique solution that attains this maximum in C. Like u_0, we can show that u_α is also a merit function in the sense of weak Pareto optimality.

Theorem 3.3
Let u_α be defined by (3.2) for some $\alpha > 0$. Then, we have $u_\alpha(x) \geq 0$ for all $x \in C$. Moreover, $x \in C$ is weakly Pareto optimal for (1.1) if and only if $u_\alpha(x) = 0$.

Proof. Let $x \in C$. The definition (3.2) of u_α yields

$$\begin{aligned} u_\alpha(x) &= \max_{y \in C} \min_{i=1,\dots,m} \left[F_i(x) - F_i(y) - \frac{\alpha}{2}\|x-y\|_2^2 \right] \\ &\geq \min_{i=1,\dots,m} \left[F_i(x) - F_i(x) - \frac{\alpha}{2}\|x-y\|_2^2 \right] = 0, \end{aligned}$$

which proves the first statement.

We now show the second statement. First, assume that $u_\alpha(x) = 0$. Then, (3.2) again gives

$$\min_{i=1,\dots,m} \left[F_i(x) - F_i(y) - \frac{\alpha}{2}\|x-y\|_2^2 \right] \quad \text{for all } y \in C.$$

Let $z \in C$ and $\alpha \in (0,1)$. Since the convexity of C implies that $x + \alpha(z-x) \in C$, substituting $y = x + \alpha(z-x)$ into the above inequality, we get

$$\min_{i=1,\dots,m} \left[F_i(x) - F_i(x+\alpha(z-x)) - \frac{\alpha}{2}\|\alpha(z-x)\|_2^2 \right] \leq 0.$$

The convexity of F_i leads to

$$\min_{i=1,\dots,m} \left[\alpha(F_i(x) - F_i(z)) - \frac{\alpha}{2}\|\alpha(z-x)\|_2^2 \right] \leq 0.$$

Dividing both sides by α and letting $\alpha \searrow 0$, we have

$$\min_{i=1,\dots,m} [F_i(x) - F_i(z)] \leq 0.$$

Since z can take an arbitrary point in C, it follows from (3.1) that $u_0(x) = 0$. Therefore, from Theorem 3.1, x is weakly Pareto optimal.

Now, suppose that x is weakly Pareto optimal. Then, it follows again from Theorem 3.1 that $u_0(x) = 0$. It is clear that $u_\alpha(x) \leq u_0(x)$ from the definitions (3.1) and (3.2) of u_0 and u_α, so we get $u_\alpha(x) = 0$. □

Let us now write

$$U_\alpha(x) \coloneqq \operatorname*{argmax}_{y \in C} \min_{i=1,\dots,m} \left[F_i(x) - F_i(y) - \frac{\alpha}{2}\|x-y\|_2^2 \right]. \tag{3.3}$$

Then, we can also show the continuity of u_α and U_α without any particular assumption.

Theorem 3.4

For all $\alpha > 0$, u_α and U_α defined by (3.2) and (3.3) are locally Lipschitz continuous

and locally Hölder continuous with exponent $1/2$ *on* C*, respectively.*

Proof. The optimality condition of the maximization problem associated with (3.2) and (3.3) gives

$$\alpha[x - U_\alpha(x)] \in \operatorname{conv}_{i\in\mathcal{I}(x)} \partial F_i(U_\alpha(x)) + N_C(U_\alpha(x)) \quad \text{for all } x \in C,$$

where N_C denotes the normal cone to the convex set C and

$$\mathcal{I}(x) = \operatorname*{argmin}_{i=1,\dots,m}[F_i(x) - F_i(U_\alpha(x))].$$

Thus, for all $x \in C$ there exists $\lambda(x) \in \Delta^m$, where Δ^m is the unit simplex given by (2.1), such that $\lambda_j(x) \neq 0$ for all $j \notin \mathcal{I}(x)$ and

$$\alpha\langle x - U_\alpha(x), z - U_\alpha(x)\rangle \leq \sum_{i=1}^m \lambda_i(x)[F_i(z) - F_i(U_\alpha(x))] \quad \text{for all } z \in C.$$

For any bounded set $\Omega \subseteq C$, let $x^1, x^2 \in \Omega$. Adding the two inequalities obtained by substituting $(x, z) = (x^1, U_\alpha(x^2))$ and $(x, z) = (x^2, U_\alpha(x^1))$ into the above inequality, we get

$$\begin{aligned}
&\alpha\big\langle U_\alpha(x^1) - U_\alpha(x^2) - (x^1 - x^2), U_\alpha(x^1) - U_\alpha(x^2)\big\rangle \\
&\quad\leq \sum_{i=1}^m \left[\lambda_i(x^2) - \lambda_i(x^1)\right]\left[F_i(U_\alpha(x^1)) - F_i(U_\alpha(x^2))\right] \\
&\quad= \sum_{i=1}^m \lambda_i(x^1)\left[F_i(x^1) - F_i(U_\alpha(x^1))\right] + \sum_{i=1}^m \lambda_i(x^2)\left[F_i(x^2) - F_i(U_\alpha(x^2))\right] \\
&\qquad+ \sum_{i=1}^m \lambda_i(x^1)\left[F_i(U_\alpha(x^2)) - F_i(x^1)\right] + \sum_{i=1}^m \lambda_i(x^2)\left[F_i(U_\alpha(x^1)) - F_i(x^2)\right].
\end{aligned}$$

Since $\lambda(x) \in \Delta^m$ and $\lambda_j(x) \neq 0$ for all $j \in \mathcal{I}(x)$, we have

$$\begin{aligned}
&\alpha\big\langle U_\alpha(x^1) - U_\alpha(x^2) - (x^1 - x^2), U_\alpha(x^1) - U_\alpha(x^2)\big\rangle \\
&\quad= \min_{i=1,\dots,m}\left[F_i(x^1) - F_i(U_\alpha(x^1))\right] + \min_{i=1,\dots,m}\left[F_i(x^2) - F_i(U_\alpha(x^2))\right] \\
&\qquad+ \sum_{i=1}^m \lambda_i(x^1)\left[F_i(U_\alpha(x^2)) - F_i(x^1)\right] + \sum_{i=1}^m \lambda_i(x^2)\left[F_i(U_\alpha(x^1)) - F_i(x^2)\right]
\end{aligned}$$

Again using the fact that $\lambda(x) \in \Delta^m$, we get

$$\begin{aligned}
&\alpha\big\langle U_\alpha(x^1) - U_\alpha(x^2) - (x^1 - x^2), U_\alpha(x^1) - U_\alpha(x^2)\big\rangle \\
&\quad\le \sum_{i=1}^m \lambda_i(x^2)\left[F_i(x^1) - F_i(U_\alpha(x^1))\right] + \sum_{i=1}^m \lambda_i(x^1)\left[F_i(x^2) - F_i(U_\alpha(x^2))\right] \\
&\qquad + \sum_{i=1}^m \lambda_i(x^1)\left[F_i(U_\alpha(x^2)) - F_i(x^1)\right] + \sum_{i=1}^m \lambda_i(x^2)\left[F_i(U_\alpha(x^1)) - F_i(x^2)\right] \\
&\quad= \sum_{i=1}^m \left[\lambda_i(x^2) - \lambda_i(x^1)\right]\left[F_i(x^1) - F_i(x^2)\right] \le 2 \max_{i=1,\dots,m} \left|F_i(x^1) - F_i(x^2)\right|.
\end{aligned}$$

Dividing by α and adding $(1/4)\|x^1 - x^2\|^2$ in both sides of the inequality, it follows that

$$\left\|U_\alpha(x^1) - U_\alpha(x^2) - \frac{1}{2}\left(x^1 - x^2\right)\right\|_2^2 \le \frac{1}{4}\left\|x^1 - x^2\right\|^2 + \frac{2}{\alpha}\max_{i=1,\dots,m}\left|F_i(x^1) - F_i(x^2)\right|.$$

Taking the square root of both sides, we obtain

$$\left\|U_\alpha(x^1) - U_\alpha(x^2) - \frac{1}{2}\left(x^1 - x^2\right)\right\|_2 \le \sqrt{\frac{1}{4}\|x^1 - x^2\|^2 + \frac{2}{\alpha}\max_{i=1,\dots,m}|F_i(x^1) - F_i(x^2)|}.$$

Then, it follows from the triangle inequality that

$$\left\|U_\alpha(x^1) - U_\alpha(x^2)\right\|_2 \le \frac{1}{2}\left\|x^1 - x^2\right\|_2 + \sqrt{\frac{1}{4}\|x^1 - x^2\|_2^2 + \frac{2}{\alpha}\max_{i=1,\dots,m}|F_i(x^1) - F_i(x^2)|}.$$

Since Lemma 2.1 implies that F_i locally Lipschitz continuous, there exists $L_i > 0$ such that

$$|F_i(x^1) - F_i(x^2)| \le L_i\|x^1 - x^2\|_2. \tag{3.4}$$

Hence, the above two inequalities show U_α's local Hölder continuity with exponent $1/2$.

On the other hand, the definition (3.2) of u_α gives

$$\begin{aligned}
u_\alpha(x^1) &= \max_{y\in C}\min_{i=1,\dots,m}\left[F_i(x^1) - F_i(y) - \frac{\alpha}{2}\|x^1 - y\|_2^2\right] \\
&\ge \min_{i=1,\dots,m}[F_i(x^1) - F_i(U_\alpha(x^2)] - \frac{\alpha}{2}\|x^1 - U_\alpha(x^2)\|_2^2.
\end{aligned}$$

Reducing $u_\alpha(x^2)$ from both sides yields

$$u_\alpha(x^1) - u_\alpha(x^2) \geq \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) - \frac{\alpha}{2}\left\|x^1 - U_\alpha(x^2)\right\|_2^2 \right] - u_\alpha(x^2).$$

Eqs. (3.2) and (3.3) lead to

$$\begin{aligned} u_\alpha(x^1) - u_\alpha(x^2) \geq & \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) - \frac{\alpha}{2}\left\|x^1 - U_\alpha(x^2)\right\|_2^2 \right] \\ & - \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) - \frac{\alpha}{2}\left\|x^2 - U_\alpha(x^2)\right\|_2^2 \right]. \end{aligned}$$

From the relation $\min_{i=1,\dots,m} v_i^1 - \min_{i=1,\dots,m} v_i^2 \geq \min_{i=1,\dots,m}(v_i^1 - v_i^2)$ for all $v^1, v^2 \in \mathbf{R}^m$, we obtain

$$u_\alpha(x^1) - u_\alpha(x^2) \geq \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(x^2) - \frac{\alpha}{2}\left\langle x^1 + x^2 - 2U_\alpha(x^2), x^1 - x^2 \right\rangle \right].$$

Cauchy-Schwarz inequality and (3.4) implies

$$u_\alpha(x^1) - u_\alpha(x^2) \geq -\left[\max_{i=1,\dots,m} L_i + \frac{\alpha}{2}\left\|x^1 + x^2 - 2U_\alpha(x^2)\right\|_2 \right] \left\|x^1 - x^2\right\|_2.$$

Since the above inquality holds even if we interchange x^1 and x^2, we can show the local Lipschitz continuity of u_α. □

On the other hand, using the unit simplex Δ^m defined by (2.1), u_α given by (3.2) can also be expressed as

$$u_\alpha(x) = \max_{y \in C} \min_{\lambda \in \Delta^m} \sum_{i=1}^m \lambda_i \left[F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \right].$$

We can see that C is convex, Δ^m is compact and convex, and the function inside $\min_{\lambda \in \Delta^m}$ is convex for λ and concave for y. Therefore, Sion's minimax theorem [**Sion1958**] leads to

$$\begin{aligned} u_\alpha(x) &= \min_{\lambda \in \Delta^m} \max_{y \in C} \sum_{i=1}^m \lambda_i \left[F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \right] \\ &= \min_{\lambda \in \Delta^m} \left[\sum_{i=1}^m \lambda_i F_i(x) - \alpha\, \mathcal{M}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i + \delta_C}(x) \right], \end{aligned} \tag{3.5}$$

where $\mathcal{M}$ and δ denote the Moreau envelope and the indicator function defined

by (1.7) and (2.6), respectively. Thus, we can evaluate u_α through the following m-dimensional, simplex-constrained, and convex optimization problem:

$$\begin{aligned} \min_{\lambda \in \mathbf{R}^m} \quad & \sum_{i=1}^m \lambda_i F_i(x) - \alpha \, \mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x) \\ \text{s.t.} \quad & \lambda \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1. \end{aligned} \tag{3.6}$$

As the following theorem shows, (3.6) is also differentiable.

Theorem 3.5

The objective function of (3.6) is continuously differentiable at every $\lambda \in \mathbf{R}^m$ and

$$\nabla_\lambda \left[\sum_{i=1}^m \lambda_i F_i(x) - \alpha \, \mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m F_i + \delta_C}(x) \right] = F(x) - F\left(\mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x) \right),$$

where **prox** *denotes the proximal operator (2.7).*

Proof. Define

$$h(y, \lambda) \coloneqq \sum_{i=1}^m \lambda_i F_i(y) + \frac{\alpha}{2} \|x - y\|_2^2.$$

Clearly, h is continuous. Moreover, $h_y(\cdot) \coloneqq h(y, \cdot)$ is continuously differentiable and

$$\nabla_\lambda h_y(\lambda) = F(y).$$

Furthermore, $\mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x) = \operatorname{argmin}_{y \in C} h(y, \lambda)$ is also continuous at every $\lambda \in \mathbf{R}^m$ from [**Rockafellar1998**]. Therefore, all the assumptions of Proposition 2.7 are satisfied. Since $\mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$ is unique, we obtain the desired result. □

Therefore, when $\mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$ is easy to compute, we can solve (3.6) using well-known convex optimization techniques such as the interior point method [**Bertsekas1999**]. If $n \gg m$, this is usually faster than solving the n-dimensional problem directly to compute (3.2).

Let us now write the optimal solution set of (3.6) by

$$\Lambda(x) \coloneqq \operatorname*{argmin}_{\lambda \in \Delta^m} \left[\sum_{i=1}^m \lambda_i F_i(x) - \alpha \, \mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x) \right]. \tag{3.7}$$

Then, we can show the directional differentiability of u_α, as in the following theorem.

Theorem 3.6
Let $x \in C$. For all $\alpha > 0$, the regularized gap function function u_α defined by (3.2) has a directional derivative

$$u'_\alpha(x; z-x) = \inf_{\lambda \in \Lambda(x)} \left[\sum_{i=1}^m \lambda_i F_i'(x; z-x) - \alpha \left\langle x - \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x), z-x \right\rangle \right]$$

for all $z \in C$, where $\Lambda(x)$ is given by (3.7), and **prox** *denotes the proximal operator (2.7). In particular, if $\Lambda(x)$ is a singleton, i.e., $\Lambda(x) = \{\lambda(x)\}$, and F_i is continuously differentiable at x, then u_α is continuously differentiable at x, and we have*

$$\nabla u_\alpha(x) = \sum_{i=1}^m \lambda_i(x) \nabla F_i(x) - \alpha \left(x - \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i(x) F_i + \delta_C} \right).$$

Proof. Let

$$h(x, \lambda) := \sum_{i=1}^m \lambda_i F_i(x) - \alpha \, \mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x).$$

Since $\mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$ is continuous at every $(x, \lambda) \in C \times \Delta^m$ from [**Rockafellar1998**], h is also continuous on $C \times \Delta^m$. Moreover, Theorem 2.4 implies that for all $x, z \in C$ the function $h_\lambda(\cdot) := h(\cdot, \lambda)$ has a directional derivative:

$$h'_\lambda(x; z-x) = \sum_{i=1}^m \lambda_i F_i'(x; z-x) - \alpha x - \left\langle \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x), z-x \right\rangle.$$

Because $\mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$ is continuous at every $(x, \lambda) \in C \times \Delta^m$ (cf. [**Rockafellar1998**]), $h'_\lambda(x; z-x)$ is also continuous at every $(x, z, \lambda) \in C \times C \times \Delta^m$. The discussion above and the compactness of Δ^m show that all assumptions of Proposition 2.7 are satisfied, so we get the desired result. □

From Theorems 3.3 and 3.6, the weakly Pareto optimal solutions for (1.1) are the globally optimal solutions of the following (directionally) differentiable single-objective optimization problem:

$$\min_{x \in C} \quad u_\alpha(x). \tag{3.8}$$

Since u_α is generally non-convex, (3.8) may have local optimal solutions or stationary points that are not globally optimal. As the following example shows, such stationary points are not necessarily Pareto stationary for (1.1).

Example 3.1
Let $m = 1, \alpha = 1, S = \mathbf{R}$ and $F_1(x) = |x|$. Then, we have

$$\mathcal{M}_{F_1}(x) = \begin{cases} x^2/2, & \text{if } |x| < 1, \\ |x| - 1/2, & \text{otherwise.} \end{cases}$$

Hence, we can evaluate u_1 as follows:

$$u_1(x) = \begin{cases} |x| - x^2/2, & \text{if } |x| < 1, \\ 1/2, & \text{otherwise.} \end{cases}$$

It is stationary for (3.8) at $|x| \geq 1$ and $x = 0$ but minimal only at $x = 0$. Furthermore, the stationary point of F_1 is only $x = 0$.

However, if we assume the strict convexity of each F_i, the stationary point of (3.8) is Pareto optimal for (1.1) and hence global optimal for (3.8). Note that this assumption does not assert the convexity of u_α.

Theorem 3.7
Suppose that F_i is strictly convex for all $i = 1, \ldots, m$. If $x \in C$ is a stationary point of (3.8), i.e.,

$$u'_\alpha(x; z - x) \geq 0 \quad \text{for all } z \in C,$$

then x is Pareto optimal for (1.1).

Proof. Let $\lambda \in \Lambda(x)$, where $\Lambda(x)$ is given by (3.7). Then, Theorem 3.6 gives

$$\sum_{i=1}^m \lambda_i F_i'(x; z - x) - \alpha \left\langle x - \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i + \delta_C}(x), z - x \right\rangle \geq 0 \quad \text{for all } z \in C.$$

Substituting $z = \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$ into the above inequality, we get

$$\sum_{i=1}^m \lambda_i F_i' \left(x; \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i + \delta_C}(x) - x \right) + \alpha \left\| x - \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i + \delta_C}(x) \right\|_2^2 \geq 0.$$

On the other hand, Theorem 2.5 yields

$$\left\| x - \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^{m}\lambda_i F_i+\delta_C}(x) \right\|_2^2 \leq \frac{1}{\alpha}\sum_{i=1}^{m}\lambda_i \left[F_i(x) - F_i\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^{m}\lambda_i F_i+\delta_C}(x) \right) \right].$$

Combining the above two inequalities, we have

$$\sum_{i=1}^{m}\lambda_i F_i'\left(x; \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^{m}\lambda_i F_i+\delta_C}(x) - x \right) \geq \sum_{i=1}^{m}\lambda_i \left[F_i\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^{m}\lambda_i F_i+\delta_C}(x) \right) - F_i(x) \right].$$

Since F_i is strictly convex for all $i = 1, \ldots, m$, the above inequality implies that $x = \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^{m}\lambda_i F_i+\delta_C}(x)$, and hence $u_\alpha(x) = 0$. This means that x is Pareto optimal for (1.1) from the strict convexity of F_i, Lemma 2.8 (i), (iii) , and Theorem 3.3.□

3.2.3 A regularized and partially linearized gap function for composite multi-objective optimization

Now, let us consider the composite case, i.e., each component F_i of the objective function F of (1.1) has the composite structure (1.13) Since they are generally non-convex, we can regard them as a relaxation of the assumptions of the previous subsection. For (1.1) with objective function (1.13), we propose a regularized and partially linearized gap function $w_\alpha \colon C \to \mathbf{R}$ with a given $\alpha > 0$ as follows:

$$w_\alpha(x) \coloneqq \max_{y\in C} \min_{i=1,\ldots,m} \left[\langle \nabla f_i(x), x - y\rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \right]. \tag{3.9}$$

Like u_α, the convexity of g_i leads to the finiteness of w_α and the existence of a unique solution that attains $\max_{y\in C}$. As the following remark shows, w_α generalizes other kinds of merit functions.

Remark 3.1

*(i) When $g_i = 0$, w_α corresponds to the regularized gap function [**Charitha2010**] for vector variational inequality.*

(ii) When $f_i = 0$, w_α matches u_α defined by (3.2).

As shown in the following theorem, w_α is a merit function in the sense of Pareto stationarity.

Theorem 3.8
Let w_α be given by (3.9) for some $\alpha > 0$. Then, we have $w_\alpha(x) \geq 0$ for all $x \in C$. Furthermore, $x \in C$ is Pareto stationary for (1.1) if and only if $w_\alpha(x) = 0$.

Proof. We first show the nonnegativity of w_α for all $\alpha > 0$. Let $x \in C$. The definition of w_α gives

$$\begin{aligned} w_\alpha(x) &= \sup_{y \in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2} \|x - y\|_2^2 \right] \\ &\geq \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - x \rangle + g_i(x) - g_i(x) - \frac{\alpha}{2} \|x - x\|_2^2 \right] = 0. \end{aligned}$$

Let us prove the second statement. Assume that $w_\alpha(x) = 0$. Then, again using the definition of w_α, we get

$$\min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2} \|x - y\|_2^2 \right] \leq 0 \quad \text{for all } y \in C.$$

Let $z \in C$ and $t \in (0, 1)$. Since $C \subseteq \mathbf{R}^n$ is convex, $x, z \in C$ implies $x + t(z - x) \in C$. Therefore, by substituting $y = x + t(z - x)$ into the above inequality, we obtain

$$\min_{i=1,\dots,m} \left[-\langle \nabla f_i(x), t(z - x) \rangle + g_i(x) - g_i(x + \alpha(z - x)) - \frac{\alpha}{2} \|t(z - x)\|^2 \right] \leq 0.$$

Dividing both sides by α yields

$$\min_{i=1,\dots,m} \left[-\langle \nabla f_i(x), z - x \rangle - \frac{g_i(x + t(z - x)) - g_i(x)}{t} - \frac{\alpha t}{2} \|z - x\|^2 \right] \leq 0.$$

By taking $\alpha \searrow 0$ and multiplying both sides by -1, we get

$$\max_{i=1,\dots,m} F_i'(x; z - x) \geq 0,$$

which means that x is Pareto stationary for (1.1).

Now, we prove the converse by contrapositive. Suppose that $w_\alpha(x) > 0$. Then, from the definition of w_α, there exists some $y \in C$ such that

$$\min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2} \|x - y\|_2^2 \right] > 0.$$

Since g_i is convex, we obtain

$$\min_{i=1,\dots,m}\left[\langle\nabla f_i(x), x-y\rangle - g_i'(x;y-x) - \frac{\alpha}{2}\|x-y\|_2^2\right] > 0.$$

Thus, we have

$$\max_{i=1,\dots,m} F_i'(x;y-x) \leq -\frac{\alpha}{2}\|x-y\|_2^2 < 0,$$

which shows that x is not Pareto stationary for (1.1). □

While u_0 and u_α given by (3.1) and (3.2) are merit functions in the sense of weak Pareto optimality, w_α defined by (3.9) is a merit function only in the sense of Pareto stationarity. As indicated by the following example, even if $w_\alpha(x) = 0$, x is not necessarily weakly Pareto optimal for (1.1).

Example 3.2
Consider the single-objective function $F\colon \mathbf{R} \to \mathbf{R}$ *defined by* $F(x) \coloneqq f(x) + g(x)$, *where*

$$f(x) \coloneqq -x^2 \quad \textit{and} \quad g(x) \coloneqq 0,$$

and set $C = \mathbf{R}$. *Then, we have*

$$w_\alpha(0) = \max_{y\in\mathbf{R}}\left[f'(0)(0-y) + g(0) - g(y) - \frac{\alpha}{2}(y-0)^2\right] = \max_{y\in\mathbf{R}}\left[-\frac{\alpha}{2}y^2\right] = 0,$$

but $x = 0$ *is not global minimal (i.e., weakly Pareto optimal) for* F.

We now define the optimal solution mapping $W_\alpha\colon C \to C$ associated with (3.9) by

$$W_\alpha(x) \coloneqq \operatorname*{argmax}_{y\in C} \min_{i=1,\dots,m}\left[\langle\nabla f_i(x), x-y\rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x-y\|_2^2\right]. \tag{3.10}$$

We can also show the continuity of w_α and W_α.

Theorem 3.9
For all $\alpha > 0$, w_α *and* W_α *defined by* (3.9) *and* (3.10) *are continuous on* C.

Proof. Since $y \mapsto -\langle\nabla f_i(x), x-y\rangle$ is convex and $\langle\nabla f_i(x), x-x\rangle = 0$, if we replace each f_i by $y \mapsto -\langle\nabla f_i(x), x-y\rangle$ in (1.1), then we can regard w_α and W_α as u_α and U_α defined by (3.2) and (3.3), respectively. Therefore, Theorem 3.4 proves this theorem. □

On the other hand, in the same way as the derivation of (3.5), Sion's minimax theorem [**Sion1958**] gives another representation of w_α for $\alpha > 0$ as follows:

$$w_\alpha(x) = \min_{\pi \in \Delta^m} \max_{y \in C} \sum_{i=1}^m \pi_i \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2} \|x - y\|_2^2 \right], \quad (3.11)$$

where Δ^m denotes the standard simplex (2.1). Moreover, simple calculations show that

$$\begin{aligned} w_\alpha(x) &= \min_{\pi \in \Delta^m} \left\{ \sum_{i=1}^m \pi_i g_i(x) + \frac{1}{2\alpha} \left\| \sum_{i=1}^m \pi_i \nabla f_i(x) \right\|_2^2 \right. \\ &\qquad \left. - \min_{y \in C} \left[\sum_{i=1}^m \pi_i g_i(y) + \frac{\alpha}{2} \left\| x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) - y \right\|_2^2 \right] \right\} \\ &= \min_{\pi \in \Delta^m} \left[\sum_{i=1}^m \pi_i g_i(x) + \frac{1}{2\alpha} \left\| \sum_{i=1}^m \pi_i \nabla f_i(x) \right\|_2^2 \right. \\ &\qquad \left. - \alpha \, \mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right], \end{aligned}$$

where $\mathcal{M}$ and δ are given by (1.7) and (2.6), respectively. In other words, we can compute w_α via the following m-dimensional, simplex-constrained, and convex optimization problem:

$$\begin{aligned} \min_{\pi \in \mathbf{R}^m} \quad & \sum_{i=1}^m \pi_i g_i(x) + \frac{1}{2\alpha} \left\| \sum_{i=1}^m \pi_i \nabla f_i(x) \right\|_2^2 \\ & - \alpha \, \mathcal{M}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \\ \text{s.t.} \quad & \pi \geq 0 \quad \text{and} \quad \sum_{i=1}^m \pi_i = 1. \end{aligned} \quad (3.12)$$

Moreover, the following theorem proves that (3.12) is differentiable.

Theorem 3.10

The objective function of (3.12) is continuously differentiable at every $\pi \in \mathbf{R}^m$ and

$$\nabla_\pi \left[\sum_{i=1}^m \pi_i g_i(x) + \frac{1}{2\alpha} \left\| \sum_{i=1}^m \pi_i \nabla f_i(x) \right\|_2^2 - \alpha \, \mathcal{M}_{\frac{1}{\alpha}\sum_{i=1}^m g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right]$$
$$= g(x) - g\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right)$$
$$- \mathcal{J}_f(x) \left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) - x \right),$$

where **prox** *is the proximal operator (2.7), and $\mathcal{J}_f(x)$ is the Jacobian matrix at x given by (2.3).*

Proof. Let

$$\theta(y, \lambda) := \sum_{i=1}^m \lambda_i g_i(y) + \frac{\alpha}{2} \left\| x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) - y \right\|_2^2.$$

Then, θ is continuous, $\theta_y(\cdot) := \theta(y, \cdot)$ is continuously differentiable, and

$$\nabla_\pi \theta_y(\pi) = g(y) + \mathcal{J}_f(x) \left(y - x + \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right).$$

Moreover, $\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C}(x) = \operatorname{argmin}_{y \in C} \theta(y, \lambda)$ is also continuous at every $\pi \in \mathbf{R}^m$ (cf. [**Rockafellar1998**]). The above discussion implies that every assumption in Proposition 2.7 is satisfied. Combined with the uniqueness of $\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C}(x)$, we get

$$\nabla_\pi \left[\alpha \, \mathcal{M}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right]$$
$$= g\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right)$$
$$+ \mathcal{J}_f(x) \left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) - x + \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right).$$

On the other hand, we have

$$\nabla_\pi \left[\sum_{i=1}^m \pi_i g_i(x) + \frac{1}{2\alpha} \left\| \sum_{i=1}^m \pi_i \nabla f_i(x) \right\|_2^2 \right] = g(x) + \frac{1}{\alpha} \mathcal{J}_f(x) \sum_{i=1}^m \pi_i \nabla f_i(x).$$

Adding the above two equalities, we obtain the desired result. □

Thus, like (3.6), (3.12) is solvable with convex optimization techniques such as the interior point method [**Bertsekas1999**] when we can quickly evaluate $\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i g_i + \delta_C}(\cdot)$. When $n \gg m$, this usually gives a faster way to compute w_α. Note, for example, that if $g_i(x) = 0$ for all $i = 1, \ldots, m$, then $\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i g_i + \delta_C}$ reduces to the projection onto C from (2.8). Moreover, for example, if $g_i(x) = g_1(x)$ for any $i = 1, \ldots, m$, or if $g_i(x) = g_1(x_{I_i})$ and the index sets $I_i \subseteq \{1, \ldots, n\}$ do not overlap each other, then $\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i g_i}$ is computable with each $\mathbf{prox}_{g_i}$ when $C = \mathbf{R}^n$.

Now, define the optimal solution set of (3.12) by

$$\Pi(x) = \operatorname*{argmin}_{\pi \in \Delta^m} \left[\sum_{i=1}^m \pi_i g_i(x) + \frac{1}{2\alpha} \left\| \sum_{i=1}^m \pi_i \nabla f_i(x) \right\|_2^2 - \alpha \mathcal{M}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right]. \quad (3.13)$$

Then, in the same manner as Theorem 3.6, we obtain the following theorem.

Theorem 3.11

Let $x \in C$. Assume that f_i is twice continuously differentiable at x. Then, for all $\alpha > 0$, the merit function w_α defined by (3.9) has a directional derivative

$$w'_\alpha(x; z - x) = \inf_{\pi \in \Pi(x)} \left[\sum_{i=1}^m \pi_i g'_i(x; z - x) - \alpha \left\langle \left[I_n - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla^2 f_i(x) \right] \left[x - \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right] - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x), z - x \right\rangle \right]$$

for all $z \in C$, where **prox** *and Π is given by (2.7) and (3.13), respectively, and $I_n \in \mathbf{R}^{n \times n}$ is the n-dimensional identity matrix. In particular, if $\Pi(x)$ is a singleton,*

i.e., $\Pi(x) = \{\pi(x)\}$*, and* g_i *is continuously differentiable at* x*, then* w_α *is continuously differentiable at* x*, and we have*

$$\nabla w_\alpha(x) = \sum_{i=1}^m \pi_i(x)\nabla F_i(x) \\ -\alpha\left[I_n - \frac{1}{\alpha}\sum_{i=1}^m \pi_i(x)\nabla^2 f_i(x)\right]\left[x - \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i(x)g_i+\delta_C}\left(x - \frac{1}{\alpha}\sum_{i=1}^m \pi_i(x)\nabla f_i(x)\right)\right].$$

If the convex part g_i is the same regardless of i, we get the following corollary without assuming the differentiability of g_i.

Corollary 3.12

Let $x \in C$ *and* $\alpha > 0$*. Assume that* f_i *is twice continuously differentiable at* x *and* $g_i = g_1$ *for all* $i = 1, \ldots, m$*, and recall that* w_α *and* **prox** *be defined by (2.7) and (3.9), respectively. If* $\Pi(x)$ *given by (3.13) is a singleton, i.e.,* $\Pi(x) = \{\pi(x)\}$*, then the function* $w_\alpha - g_1$ *is continuously differentiable at* x*, and we have*

$$\nabla_x\left(w_\alpha(x) - g_1(x)\right) \\ = -\alpha\left[I_n - \frac{1}{\alpha}\sum_{i=1}^m \pi_i(x)\nabla^2 f_i(x)\right]\left[x - \mathbf{prox}_{\frac{1}{\alpha}g_1+\delta_C}\left(x - \frac{1}{\alpha}\sum_{i=1}^m \pi_i(x)\nabla f_i(x)\right)\right] \\ + \sum_{i=1}^m \pi_i(x)\nabla f_i(x).$$

Corollary 3.12 implies that, under certain conditions, the merit function $w_\alpha = (w_\alpha - g_1) + g_1$ is composite, i.e., the sum of a continuously differentiable function and a convex one.

Theorems 3.8 and 3.11 show that the Pareto stationary points for (1.1) are global optimal for the following directionally differentiable single-objective optimization problem:

$$\min_{x\in C} \quad w_\alpha(x). \tag{3.14}$$

Moreover, when the assumptions of Corollary 3.12 hold, we can apply first-order methods such as the proximal gradient method [**Fukushima1981**] to (3.14). On the other hand, if we consider Example 3.1 with $f_i = 0$, we can see that the stationary point for (3.14) is not necessarily Pareto stationary for (1.1). However, if f_i is convex and twice continuously differentiable, and F_i is strictly convex, then we can prove

that every stationary point of (3.14) is Pareto optimal for (1.1), i.e., global optimal for (3.8). Note that this assumption does not assert the convexity of w_α.

Theorem 3.13
Let $x \in C$ and $\alpha > 0$. Suppose that f_i is convex and twice continuously differentiable at x, and F_i is strictly convex for any $i = 1, \ldots, m$. If x is stationary for (3.14), i.e.,

$$w'_\alpha(x; z - x) \geq 0 \quad \textit{for all } z \in C,$$

then x is Pareto optimal for (1.1).

Proof. Let $z \in C$ and $\pi \in \Pi(x)$, where $\Pi(x)$ is defined by (3.13). Then, it follows from Theorem 3.11 that

$$\begin{aligned}
&\sum_{i=1}^m \pi_i g'_i(x; z - x) \\
&\quad - \alpha \left\langle \left[I_n - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla^2 f_i(x) \right] \left[x - \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right] \right. \\
&\qquad \left. - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x), z - x \right\rangle \geq 0.
\end{aligned}$$

Substituting $z = \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$, we have

$$\begin{aligned}
&\sum_{i=1}^m \pi_i F'_i \left(x; \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) - x \right) \\
&\quad + \alpha \left\langle \left[I_n - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla^2 f_i(x) \right] \left[x - \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right] , \right. \\
&\qquad \left. x - \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right\rangle \geq 0.
\end{aligned}$$

Since the convexity of f_i implies that $\nabla^2 f_i(x)$ is positive semidefinite, we get

$$\begin{aligned}
&\sum_{i=1}^m \pi_i F'_i \left(x; \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) - x \right) \\
&\qquad + \alpha \left\| \mathbf{prox}_{\frac{1}{\alpha} \sum_{i=1}^m \pi_i g_i + \delta_C} \left(x - \frac{1}{\alpha} \sum_{i=1}^m \pi_i \nabla f_i(x) \right) \right\|_2^2 \geq 0.
\end{aligned}$$

Therefore, with similar arguments used in the proof of Theorem 3.7, we obtain $x = \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \pi_i g_i + \delta_C}(x - (1/\alpha)\sum_{i=1}^m \pi_i \nabla f_i(x))$, and thus $w_\alpha(x) = 0$. Since F_i is strictly convex, x is Pareto optimal for (1.1) from Lemma 2.8 (iii) and Theorem 3.8. □

3.3 Relation between different merit functions

This section assumes that the problem has a composite structure (1.13) and discusses the connection between the merit functions proposed in Sections 3.2.1–3.2.3. First, we show some inequalities between different types of merit functions.

Theorem 3.14
Let u_0, u_α, and w_α be defined by (3.1), (3.2) and (3.9), respectively, for all $\alpha > 0$. Then, the following statements hold.

(i) If f_i is μ_i-convex for some $\mu_i \in \mathbf{R}$ and $\mu = \min_{i=1,\dots,m} \mu_i$, then we have

$$\begin{cases} u_0(x) \le w_\mu(x) \quad \text{and} \quad u_\alpha(x) \le w_{\mu+\alpha}(x), & \text{if } \mu \ge 0, \\ u_{-\mu+\alpha}(x) \le w_\alpha(x), & \text{otherwise} \end{cases}$$

for all $\alpha > 0$ and $x \in C$.

(ii) If ∇f_i is L_i-Lipschitz continuous for some $L_i > 0$ and $L = \max_{i=1,\dots,m} L_i$, then we get

$$u_{L+\alpha}(x) \le w_\alpha(x), \quad u_0(x) \ge w_L(x), \quad \text{and} \quad u_\alpha(x) \ge w_{L+\alpha}(x)$$

for all $\alpha > 0$ and $x \in C$.

Proof. Claim (i) : Let $i = 1, \dots, m$. The μ_i-convexity of f_i gives

$$f_i(x) - f_i(y) \le \langle \nabla f_i(x), x - y \rangle - \frac{\mu_i}{2}\|x - y\|_2^2.$$

By the definition of μ, we get

$$f_i(x) - f_i(y) \le \langle \nabla f_i(x), x - y \rangle - \frac{\mu}{2}\|x - y\|_2^2.$$

Thus, recalling (1.13), we have

$$F_i(x) - F_i(y) \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\mu}{2}\|x - y\|^2,$$
$$F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\mu + \alpha}{2}\|x - y\|^2,$$
$$F_i(x) - F_i(y) - \frac{-\mu + \alpha}{2}\|x - y\|_2^2 \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x - y\|^2,$$

so the desired inequalities are clear from (3.1), (3.2) and (3.9).

Claim (ii) : Let $i = 1, \ldots, m$. Suppose that ∇f_i is L_i-Lipschitz continuous. Then, Lemma 2.2 yields

$$\left| f_i(y) - f_i(x) - \nabla f_i(x)^\top (y - x) \right| \leq \frac{L_i}{2}\|x - y\|_2^2.$$

By the definition of L, we have

$$\left| f_i(y) - f_i(x) - \nabla f_i(x)^\top (y - x) \right| \leq \frac{L}{2}\|x - y\|_2^2.$$

This gives

$$F_i(x) - F_i(y) - \frac{L + \alpha}{2}\|x - y\|_2^2 \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x - y\|_2^2,$$
$$F_i(x) - F_i(y) \geq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{L}{2}\|x - y\|_2^2,$$
$$F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \geq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{L + \alpha}{2}\|x - y\|_2^2.$$

Therefore, we immediately get $u_{L+\alpha}(x) \leq w_\alpha(x)$, $u_0(x) \geq w_L(x)$, and $u_\alpha(x) \geq w_{L+\alpha}(x)$ for all $x \in C$ by (3.1), (3.2) and (3.9). □

Second, we present the relation between coefficients and the proposed merit functions' values.

Theorem 3.15

Recall that w_α is defined by (3.9) for all $\alpha > 0$. Let r be an arbitrary scalar such that $r \geq \alpha$. Then, we get

$$w_r(x) \leq w_\alpha(x) \leq \frac{r}{\alpha} w_r(x) \quad \textit{for all } x \in C.$$

Proof. Let $x \in C$. Since $r \geq \alpha > 0$, the definition (3.9) of w_r and w_α clearly gives the first inequality. Thus, we prove the second one. From the definition (3.9) of w_α, we have

$$\begin{aligned} w_\alpha(x) &= \sup_{y\in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x-y\rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x-y\|_2^2\right] \\ &= \frac{r}{\alpha} \sup_{y\in C} \min_{i=1,\dots,m} \left[\left\langle \nabla f_i(x), \frac{\alpha}{r}(x-y)\right\rangle + \frac{\alpha}{r}(g_i(x)-g_i(y)) - \frac{r}{2}\left\|\frac{\alpha}{r}(x-y)\right\|_2^2\right] \\ &\leq \frac{r}{\alpha} \sup_{y\in C} \min_{i=1,\dots,m} \left[\left\langle \nabla f_i(x), \frac{\alpha}{r}(x-y)\right\rangle + g_i(x) - g_i\left(x - \frac{\alpha}{r}(x-y)\right)\right. \\ &\qquad \left. - \frac{r}{2}\left\|\frac{\alpha}{r}(x-y)\right\|_2^2\right] \end{aligned}$$

where the first inequality follows from the convexity of g_i. Since C is convex, $x, y \in C$ implies $x - (\alpha/r)(x-y) \in C$. Therefore, from the definition (3.9) of w_r, we get

$$w_\alpha(x) \leq \frac{r}{\alpha} w_r(x). \qquad \square$$

Considering Remark 3.1 (ii), we get the following corollary.

Corollary 3.16
Assume that each component F_i of the objective function F of (1.1) is convex. Recall that u_α is defined by (3.2) for all $\alpha > 0$. Let r be an arbitrary scalar such that $r \geq \alpha$. Then, we get

$$u_r(x) \leq u_\alpha(x) \leq \frac{r}{\alpha} u_r(x) \quad \text{for all } x \in C.$$

Remark 3.2
For unconstrained problems, we can consider the following inequality.

$$w_L(x) \geq \tau u_0(x) \quad \text{for all } x \in \mathbf{R}^n \text{ for some } \tau > 0, \tag{3.15}$$

*which is an extension of the proximal-PL condition for scalar optimization [**Karimi2016**]. Under this condition, we can prove that proximal gradient methods for multi-objective optimization [**Tanabe2019**] have linear convergence rate [**Tanabe2022**]. Note that this inequality holds particularly if each f_i is strongly convex from Theorem 3.14 (i) and Theorem 3.15.*

3.4 Level-boundedness of the proposed merit functions

Recall that we call a function *level-bounded* if every level set is bounded. This is an important property because it ensures that the sequences generated by descent methods have accumulation points. We state below sufficient conditions for the level-boundedness of the merit functions proposed in Chapter 3.

Theorem 3.17
Let u_0, u_α and w_α be defined by (3.1), (3.2) and (3.9), respectively, for all $\alpha > 0$. Then, the following claims hold.

(i) If F_i is level-bounded for all $i = 1, \ldots, m$, then u_0 is level-bounded.

(ii) If F_i is convex and level-bounded for all $i = 1, \ldots, m$, then u_α is level-bounded for all $\alpha > 0$.

(iii) Suppose that F has the composite structure (1.13). If f_i is μ_i-convex for some $\mu_i \in \mathbf{R}$ or ∇f_i is L_i-Lipschitz continuous for some $L_i > 0$, and F_i is convex and level-bounded for all $i = 1, \ldots, m$, then w_α is level-bounded for all $\alpha > 0$.

Proof. Claim (i) : Suppose, contrary to our claim, that u_0 is not level-bounded. Then, there exists $\alpha \in \mathbf{R}$ such that $\{x \in C \mid u_0(x) \leq \alpha\}$ is unbounded. By the definition (3.2) of u_0, the inequality $u_0(x) \leq \alpha$ can be written as

$$\sup_{y \in C} \min_{i=1,\ldots,m} [F_i(x) - F_i(y)] \leq \alpha.$$

This implies that for some fixed $z \in C$, there exists $j = 1, \ldots, m$ such that

$$F_j(x) \leq F_j(z) + \alpha.$$

Therefore, it follows that

$$\{x \in C \mid u_0(x) \leq \alpha\} \subseteq \bigcup_{j=1}^{m} \{x \in C \mid F_j(x) \leq F_j(z) + \alpha\}.$$

Since F_i is level-bounded for all $i = 1, \ldots, m$, the right-hand side must be bounded, which contradicts the unboundedness of the left-hand side.

Claim (ii) : Recall the definitions (2.1), (2.6) and (2.7) of $\Delta^m, \mathcal{M}$, and **prox**. Eq. (3.5) gives

$$\begin{aligned} u_\alpha(x) &= \min_{\lambda\in\Delta^m} \left[\sum_{i=1}^m \lambda_i F_i(x) - \alpha\, \mathcal{M}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i + \delta_S}(x)\right] \\ &= \min_{\lambda\in\Delta^m} \sum_{i=1}^m \lambda_i \left[F_i(x) - F_i\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i+\delta_S}(x)\right)\right. \\ &\qquad\qquad \left. -\frac{\alpha}{2}\left\|x - \mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i+\delta_S}(x)\right\|_2^2\right] \\ &\geq \frac{1}{2}\min_{\lambda\in\Delta^m} \sum_{i=1}^m \lambda_i \left[F_i(x) - F_i\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i+\delta_S}(x)\right)\right] \\ &= \frac{1}{2}\min_{i=1,\ldots,m} \left[F_i(x) - F_i\left(\mathbf{prox}_{\frac{1}{\alpha}\sum_{i=1}^m \lambda_i F_i+\delta_S}(x)\right)\right], \end{aligned}$$

where the inequality follows from Theorem 2.5. Therefore, with similar arguments given in the proof of claim (i) , we can show the level-boundedness of u_α by contradiction.

Claim (iii) : From Theorems 3.14 and 3.15, there exist some $a > 0$ and $r > 0$ such that $u_r(x) \leq a w_\alpha(x)$ for all $x \in C$. Since claim (ii) implies that u_r is level-bounded, w_α is also level-bounded. □

As indicated by the following example, our proposed merit functions are not necessarily level-bounded even if F is level-bounded.

Example 3.3
Consider the bi-objective function $F\colon \mathbf{R} \to \mathbf{R}^2$ *with each component given by*

$$F_1(x) \coloneqq x^2, \quad F_2(x) \coloneqq 0.$$

Then, the merit function u_0 *defined by* (3.2) *is written as*

$$\begin{aligned} u_0(x) &= \sup_{y\in\mathbf{R}} \min[F_1(x) - F_1(y), F_2(x) - F_2(y)] \\ &= \sup_{y\in\mathbf{R}} \min[(x^2 - y^2), 0] = 0. \end{aligned}$$

On the other hand, F *is level-bounded because* $\lim_{\|x\|_2\to\infty} F_1(x) = \infty$.

3.5 Error bounds of the proposed merit functions

This section shows that our proposed merit functions provide global error bounds for (1.1), i.e., for any given point x, they can bound the distance between x and the Pareto optimal solution set at x multiplied by some positive constant.

Theorem 3.18

For an arbitrary $\alpha > 0$, let w_α be defined by (3.9). Assume that F has the composite structure (1.13), and at least one of the following conditions holds for all $i = 1, \ldots, m$:

(i) f_i is μ_i-convex for some $\mu_i \in \mathbf{R}$, F_i is σ_i-convex for some $\sigma_i > 0$, and $\rho_i := \sigma_i + \mu_i > 0$;

(ii) ∇f_i is L_i-Lipschitz continuous for some $L_i > 0$, F_i is σ_i-convex for some $\sigma_i > 0$, and $\rho_i := \sigma_i - L_i > 0$.

Then, we have

$$w_\alpha(x) \geq \kappa(\rho) \inf_{x^* \in X^*} \|x - x^*\|_2^2 \quad \text{for all } x \in C,$$

where X^* is the set of Pareto optimal solutions for (1.1), $\rho := \min_{i=1,\ldots,n} \rho_i$, and

$$\kappa(\rho) := \begin{cases} \dfrac{\rho - \alpha}{2}, & \text{if } \alpha < \rho/2, \\ \dfrac{\rho^2}{8\alpha}, & \text{otherwise.} \end{cases}$$

Proof. Let $x \in C$ and $r := \min(\alpha, \rho/2)$. Since $r \leq \alpha$ and $\kappa(\rho) = (r/\alpha)(\rho - r)/2$, from Theorem 3.15, it suffices to show that

$$w_r(x) \geq \frac{\rho - r}{2} \inf_{x^* \in X^*} \|x - x^*\|_2^2. \tag{3.16}$$

Recall the min-max reformulation (3.11) of w_r given by

$$w_r(x) = \min_{\gamma \in \Delta^m} \max_{y \in C} \sum_{i=1}^m \gamma_i \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{r}{2} \|x - y\|_2^2 \right],$$

where Δ^m denotes the standard simplex (2.1). Since Δ^m is compact, there exists $\gamma^* \in \Delta^m$ such that

$$w_r(x) = \max_{y \in C} \sum_{i=1}^m \gamma_i^* \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{r}{2} \|x - y\|_2^2 \right].$$

Let $x^* \in C$ be the solution of

$$\min_{x \in C} \quad \sum_{i=1}^{m} \gamma_i^* F_i(x). \tag{3.17}$$

Note that x^* is unique since F_i is strongly convex. Then, we obtain

$$\begin{aligned} w_r(x) &\geq \sum_{i=1}^{m} \gamma_i^* \left\{ \nabla f_i(x)^\top (x - x^*) + g_i(x) - g_i(x^*) - \frac{r}{2} \|x - x^*\|_2^2 \right\} \\ &\geq \sum_{i=1}^{m} \gamma_i^* \left\{ F_i(x) - F_i(x^*) + \frac{\rho_i - \sigma_i - r}{2} \|x - x^*\|_2^2 \right\} \\ &\geq \sum_{i=1}^{m} \gamma_i^* \frac{\rho_i - r}{2} \|x - x^*\|_2^2 \geq \frac{\rho - r}{2} \|x - x^*\|_2^2, \end{aligned}$$

where the second inequality follows from the definition of ρ_i and μ_i-convexity of f_i or L_i-Lipschitz continuity of ∇f_i, the third one comes from the σ_i-convexity of F_i, and the fourth one is obtained by the definition of ρ and the fact that $\gamma^* \in \Delta^m$. Since (3.17) is a weighting scalarization of (1.1), x^* is Pareto optimal for (1.1) [**Miettinen1998**], i.e., $x^* \in X^*$. Therefore, we get (3.16), which completes the proof. □

We can easily introduce error bound property of u_α by setting $f_i = 0$ in Theorem 3.18.

Corollary 3.19
For an arbitrary $\alpha \geq 0$, let u_α be defined by (3.2). Assume that F_i is σ_i-convex with $\sigma_i > 0$ for all $i = 1, \ldots, m$. Then, we have

$$u_\alpha(x) \geq \upsilon(\sigma) \inf_{x^* \in X^*} \|x - x^*\|_2^2 \quad \textit{for all } x \in C,$$

where X^* is the set of Pareto optimal solutions for (1.1), $\sigma \coloneqq \min_{i=1,\ldots,m} \sigma_i$, and

$$\upsilon(\sigma) \coloneqq \begin{cases} \dfrac{\sigma - \alpha}{2}, & \textit{if } \alpha < \sigma/2, \\ \dfrac{\sigma^2}{8\alpha}, & \textit{otherwise.} \end{cases}$$

When F_i is strongly convex, the min-max reformulation similar to (3.11) is also applicable to u_0, so we can likewise derive the error bounds for u_0 by setting $f_i = 0$ and $r = 0$ after (3.16) in the proof of Theorem 3.18.

Corollary 3.20
Let u_0 be defined by (3.1). Assume that F_i is σ_i-convex with $\sigma_i > 0$ for all $i = 1, \ldots, m$. Then, we have

$$u_0(x) \geq \frac{\sigma}{2} \inf_{x^* \in X^*} \|x - x^*\|_2^2 \quad \text{for all } x \in C,$$

where X^* is the set of Pareto optimal solutions for (1.1), and $\sigma := \min_{i=1,\ldots,m} \sigma_i$.

3.6 Conclusions

We first proposed a simple merit function for (1.1) in the sense of weak Pareto optimality and showed its lower semicontinuity. We also defined a regularized merit function when F is convex and discussed its continuity, the way of evaluating it, its differentiability, and the properties of its stationary points. Furthermore, when each F_i is composite, we introduced a regularized and partially linearized merit function in the sense of Pareto stationarity and showed similar properties. In addition, we gave sufficient conditions for the proposed merit functions to be level-bounded and to provide error bounds.

We can consider a natural extension of our proposed merit functions for vector problems with an infinite number of objective functions. We can also regard the generalization of other merit functions for scalar problems, such as the implicit Lagrangian [**Mangasarian1993**] and the squared Fischer-Burmeister function [**Kanzow1996**], to multi-objective and vector problems. These will be some subjects for future works.

Chapter 4

A proximal gradient method for multi-objective optimization

4.1 Introduction

This chapter proposes the proximal gradient method for the unconstrained composite multi-objective optimization, i.e., (1.1) with (1.13) and $C = \mathbf{R}^n$. The proposed method generalizes Algorithm 1.1. As it can be seen in the single objective function cases [**Beck2009**, **Tseng2010**], this method is shown to be efficient when g_i has some special structure. Moreover, we analyze the proposed method's convergence rate using the gap function (3.1) to measure the complexity.

We also observe that the problem and the proposed methods have many applications. For example, when g_i is an indicator function of a convex set S, (1.1) is equivalent to the optimization problems with constraints $x \in S$. Also, as it can be seen in Section 4.4, we can deal with robust optimization problems. These problems include uncertain parameters and basically consists in optimizing under the worst scenario. Although the literature about robust optimization is vast, the studies about robust multi-objective optimization is relatively new [**Ehrgott2014**, **Fliege2014**, **Morishita2016**].

The outline of this chapter is as follows. In Section 4.2, we propose the proximal gradient methods for unconstrained multi-objective optimization. We estimate the global convergence rates of the proposed method in Section 4.3. In Section 4.4, we apply the proposed method to robust optimization. Finally, we report some numerical experiments by solving robust multi-objective optimization problems in 4.5.

4.2 The algorithm

For given $x \in \operatorname{dom} F$ and $\ell > 0$, we consider the following minimization problem:

$$\min_{z \in \mathbf{R}^n} \quad \varphi_\ell(z; x), \tag{4.1}$$

where

$$\varphi_\ell(z; x) := \max_{i=1,\dots,m} [\langle \nabla f_i(x), z - x \rangle + g_i(z) - g_i(x)] + \frac{\ell}{2} \|z - x\|_2^2.$$

The convexity of g_i implies that $z \mapsto \varphi_\ell(z; x)$ is strongly convex, so the problem (4.1) always has a unique solution. Let us write such a solution as $p_\ell(x)$ and let $\theta_\ell(x)$ be its optimal function value, i.e.,

$$p_\ell(x) := \operatorname*{argmin}_{z \in \mathbf{R}^n} \varphi_\ell(z; x) \quad \text{and} \quad \theta_\ell(x) := \min_{z \in \mathbf{R}^n} \varphi_\ell(z; x). \tag{4.2}$$

The following proposition shows that $p_\ell(x)$ and $\theta_\ell(x)$ helps to characterize the Pareto stationarity of (1.1).

Lemma 4.1 ([Tanabe2019])
Let p_ℓ and θ_ℓ be defined in (4.2). Then, the following claims hold.

(i) The following three conditions are equivalent: (i) $x \in \mathbf{R}^n$ is Pareto stationary for (1.1); (ii) $p_\ell(x) = x$; (iii) $\theta_\ell(x) = x$.

(ii) The mappings p_ℓ and θ_ℓ are continuous.

From Lemma 4.1, we can treat $\|p_\ell(x) - x\|_\infty < \varepsilon$ for some $\varepsilon > 0$ as a stopping criteria. Now, we state below the proximal gradient method for (1.1).

Algorithm 4.1 Proximal gradient method for multi-objective optimization

Input: $x^0 \in \operatorname{dom} F$, $\ell > L/2$, $\varepsilon > 0$
Output: x^*: A weakly Pareto optimal point

1: $k \leftarrow 0$
2: **while** $\left\|p_\ell(x^k) - x^k\right\|_\infty \geq \varepsilon$ **do**
3: $\quad x^{k+1} \leftarrow p_\ell(x^k)$
4: $\quad k \leftarrow k + 1$
5: **end while**

4.3 Convergence rates analysis

Let us now analyze the convergence rates of Algorithm 4.1. Define $\psi_x \colon \mathbf{R}^n \to \mathbf{R}$ by

$$\psi_x(d) := \max_{i=1,\dots,m} \left[\nabla f_i(x)^\top d + g_i(x+d) - g_i(x)\right]. \tag{4.3}$$

Looking at Algorithm 4.1 differently, it generates a sequence $\{x^k\}$ iteratively with the following procedure:

$$x^{k+1} := x^k + d^k,$$

where d^k is a search direction. At every iteration k, we define this d^k by solving

$$d^k := \operatorname*{argmin}_{d \in \mathbf{R}^n} \left[\psi_{x^k}(d) + \frac{\ell}{2}\|d\|^2\right], \tag{4.4}$$

with a positive constant $\ell > 0$. Note that we have

$$\psi_{x^k}(d^k) + \frac{\ell}{2}\left\|d^k\right\|^2 = -w_\ell(x^k), \tag{4.5}$$

where w_ℓ is defined by (3.9). We suppose that the algorithm generates an infinite sequence of iterates from now on. The following result shows an important property for ψ_x.

Lemma 4.2 ([Tanabe2019])
Let $\{d^k\}$ be generated by Algorithm 4.1 and recall the definition (4.3) of ψ_x. Then, we have

$$\psi_{x^k}(d^k) \le -\ell\left\|d^k\right\|^2 \quad \textit{for all } k.$$

If $\ell > L$, from Lemma 2.2, for all $i = 1, \dots, m$ we have

$$F_i(x^{k+1}) - F_i(x^k) \le \langle \nabla f_i(x^k), d^k \rangle + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\left\|d^k\right\|. \tag{4.6}$$

The right-hand side of (4.6) is less than zero since d^k is the optimal solution of (4.4), so we get

$$F_i(x^{k+1}) \le F_i(x^k). \tag{4.7}$$

Remark 4.1
When the Lipschitz constant L is unknown or incomputable, we can use ℓ for (4.4) calculated by backtracking instead, i.e., we can set the initial value of ℓ appropriately

and multiply ℓ by a prespecified scalar $\gamma > 1$ at each iteration until (4.6) is satisfied. Since L is finite, the backtracking only requires a finite number of steps.

4.3.1 The non-convex case

When $m = 1$ and F_1 is not convex for (1.1), $\{\|x^k - \mathbf{prox}_{g_1/L_1}(x^k - \nabla f_1(x^k)/L_1)\|\}$ converges to zero with rate $O(\sqrt{1/k})$ if the proximal gradient method is applied, where the proximal operator "**prox**" is defined by (2.7) [**Beck2017**]. Note that when $m = 1$ we have $w_{L_1}(x^k) \geq (L_1/2)\|x^k - \mathbf{prox}_{g_1/L_1}(x^k - \nabla f_1(x^k)/L_1)\|^2$. Now, in the multi-objective context, we show below that $\{\sqrt{w_1(x^k)}\}$ still converges to zero with rate $O(\sqrt{1/k})$ with Algorithm 4.1.

Theorem 4.3

Suppose that there exists some nonempty set $\mathcal{J} \subseteq \{i = 1, \ldots, m\}$ such that if $i \in \mathcal{J}$ then $F_i(x)$ has a lower bound $F_i^{\min}$ for all $x \in \mathbf{R}^n$. Let $F^{\min} := \min_{i\in\mathcal{J}} F_i^{\min}$ and $F_0^{\max} := \max_{i=1,\ldots,m} F_i(x^0)$. Then, Algorithm 4.1 generates a sequence $\{x^k\}$ such that

$$\min_{0\leq j\leq k-1} w_1(x^j) \leq \frac{(F_0^{\max} - F^{\min})\max\{1, \ell\}}{k}.$$

Proof. Let $i \in \mathcal{J}$. From (4.6), we have

$$\begin{aligned} F_i(x^{k+1}) - F_i(x^k) &\leq \langle \nabla f_i(x^k), d^k\rangle + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\|d^k\|^2 \\ &\leq \max_{i=1,\ldots,m}\left\{\langle \nabla f_i(x^k), d^k\rangle + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\|d^k\|^2\right\} = -w_\ell(x^k), \end{aligned}$$

where the equality follows from (4.3) and (4.5). Adding up the above inequality from $k = 0$ to $k = \tilde{k} - 1$ yields that

$$F_i(x^{\tilde{k}}) - F_i(x^0) \leq -\sum_{k=0}^{\tilde{k}-1} w_\ell(x^k) \leq -\tilde{k} \min_{0\leq k\leq \tilde{k}-1} w_\ell(x^k).$$

From the definitions of $F^{\min}$ and $F_0^{\max}$, we obtain

$$\min_{0\leq k\leq \tilde{k}-1} w_\ell(x^k) \leq \frac{F_0^{\max} - F^{\min}}{\tilde{k}}.$$

Finally, from Theorem 3.15, we get

$$\min_{0\leq k\leq \tilde{k}-1} w_1(x^k) \leq \frac{(F_0^{\max} - F^{\min})\max\{1,\ell\}}{\tilde{k}}. \qquad \square$$

Remark 4.2

*When $g_i = 0$ for all i, references [***Calderon2020***, ***Fliege2019***, ***Grapiglia2015***] present the convergence rate of various multi-objective optimization methods. However, as we have mentioned in the introduction, they all evaluate the convergence rate with measures that depend on the subproblems or variables used in their algorithms. This means that the comparison in terms of complexity between different methods is not easy by using those measures. However, Theorem 4.3 analyzes the convergence rate using the merit function w_1, which can be defined uniformly by (3.9) for multi-objective optimization problems with a structure like (1.13).*

4.3.2 The convex case

For (1.1) with $m = 1$, if f_1 is convex, then $\{F_1(x^k) - F_1^*\}$ converges to zero with rate $O(1/k)$ using the proximal gradient method, where F_1^* is the optimal objective value of F_1 [**Beck2009**]. In this subsection, we show how fast $\{u_0(x^k)\}$ converges to zero with Algorithm 4.1. Let us start by proving the following lemma. Note, however, that we state it with f_i and g_i having general (nonnegative) convexity parameters[1], but they turn out to be zero in this subsection.

Lemma 4.4

Let f_i and g_i have convexity parameters $\mu_i \in \mathbf{R}$ and $\nu_i \in \mathbf{R}$, respectively, and write $\mu := \min_{i=1,\dots,m} \mu_i$ and $\nu := \min_{i=1,\dots,m} \nu_i$. Then, for all $x \in \mathbf{R}^n$ it follows that

$$\sum_{i=1}^{m} \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \leq \frac{\ell}{2}\left(\left\|x^k - x\right\|_2^2 - \left\|x^{k+1} - x\right\|_2^2\right) - \frac{\nu}{2}\left\|x^{k+1} - x\right\|_2^2 - \frac{\mu}{2}\left\|x^k - x\right\|_2^2,$$

where λ_i^k satisfies the following conditions:

[1] We say that $h\colon \mathbf{R}^n \to (-\infty, +\infty]$ has a *convexity parameter* $\varsigma \in \mathbf{R}$ if $h(\alpha x + (1-\alpha)y) \leq \alpha h(x) + (1-\alpha)h(y) - (1/2)\alpha(1-\alpha)\varsigma\|x-y\|^2$ holds for all $x, y \in \mathbf{R}^n$ and $\alpha \in [0,1]$. When $\varsigma > 0$, we call h strongly convex. Note that this definition allows non-convex cases, i.e., $\varsigma < 0$ can also be considered.

(i) *There exists* $\eta_i^k \in \partial g_i(x^k + d^k)$ *such that* $\sum_{i=1}^m \lambda_i^k(\nabla f_i(x^k) + \eta_i^k) + \ell d^k = 0$,

(ii) $\sum_{i=1}^m \lambda_i^k = 1, \lambda_i^k \geq 0 \left(i \in \mathcal{I}_{x^k}(d^k)\right)$ *and* $\lambda_i^k = 0 \left(i \notin \mathcal{I}_{x^k}(d^k)\right)$,

where $\mathcal{I}_x(d) \coloneqq \left\{i = 1, \ldots, m \;\middle|\; \psi_x(d) = \nabla f_i(x)^\top d + g_i(x+d) - g_i(x)\right\}$.

Proof. From (4.6), we have

$$F_i(x^{k+1}) - F_i(x^k) \leq \nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\left\|d^k\right\|_2^2.$$

The above inequality and convexity of f_i with modulus μ_i give

$$\begin{aligned}
&F_i(x^{k+1}) - F_i(x) = (F_i(x^k) - F_i(x)) + (F_i(x^{k+1}) - F_i(x^k)) \\
&\leq \left(\nabla f_i(x^k)^\top (x^k - x) - \frac{\mu_i}{2}\left\|x^k - x\right\|_2^2 + g_i(x^k) - g_i(x)\right) \\
&\quad + \left(\nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\left\|x^{k+1} - x^k\right\|_2^2\right) \\
&\leq \nabla f_i(x^k)^\top (x^k + d^k - x) + g_i(x^k + d^k) - g_i(x) - \frac{\mu}{2}\left\|x^k - x\right\|_2^2 + \frac{\ell}{2}\left\|d^k\right\|_2^2 \\
&\leq (\nabla f_i(x^k) + \eta_i^k)^\top (x^k + d^k - x) - \frac{\mu}{2}\left\|x^k - x\right\|_2^2 - \frac{\nu}{2}\left\|x^{k+1} - x\right\|_2^2 + \frac{\ell}{2}\left\|d^k\right\|_2^2,
\end{aligned}$$

where the second inequality follows from the definition of μ and the fact that $x^{k+1} = x^k + d^k$, and the last one comes from the convexity of g_i. Multiplying the above inequality by λ_i^k and summing for all $i = 1, \ldots, m$, the conditions (i) and (ii) give

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \\
&= -\ell (d^k)^\top (x^k + d^k - x) - \frac{\mu}{2}\left\|x^k - x\right\|_2^2 - \frac{\nu}{2}\left\|x^{k+1} - x\right\|_2^2 + \frac{\ell}{2}\left\|d^k\right\|_2^2 \\
&= -\frac{\ell}{2}\left(2(d^k)^\top (x^k - x) + \left\|d^k\right\|_2^2\right) - \frac{\mu}{2}\left\|x^k - x\right\|_2^2 - \frac{\nu}{2}\left\|x^{k+1} - x\right\|_2^2 \\
&= \frac{\ell}{2}\left(\left\|x^k - x\right\|_2^2 - \left\|x^{k+1} - x\right\|_2^2\right) - \frac{\mu}{2}\left\|x^k - x\right\|_2^2 - \frac{\nu}{2}\left\|x^{k+1} - x\right\|_2^2.
\end{aligned}$$

□

Now, we show that $\{u_0(x^k)\}$ converges to zero with rate $O(1/k)$ with Algorithm 4.1 under the following assumption.

Assumption 4.1

Let X^* *be the set of weakly Pareto optimal points for* (1.1)*, and define the level set of* F *for* $\alpha \in \mathbf{R}^m$ *by* $\Omega_F(\alpha) \coloneqq \{x \in S \mid F(x) \leq \alpha\}$*. Then, for all* $x \in \Omega_F(F(x^0))$

there exists $x^* \in X^*$ *such that* $F(x^*) \leq F(x)$ *and*

$$R := \sup_{F^* \in F(X^* \cap \Omega_F(F(x^0)))} \inf_{x \in F^{-1}(\{F^*\})} \left\|x - x^0\right\|_2^2 < \infty. \tag{4.8}$$

Remark 4.3

(i) In single-objective cases, Assumption 4.1 is valid if the optimization problem has at least one optimal solution; when $m = 1$, X^* *coincides with the optimal solution set, and the equality* $X^* \cap \Omega_F(F(x^0)) = X^*$ *holds, so we have* $R = \inf_{x \in X^*} \|x - x^0\|_2^2 < \infty$.

(ii) When the level set $\Omega_F(F(x^0))$ *is bounded, Assumption 4.1 is also satisfied. For example, this is the case when* F_i *is strongly convex for at least one* i.

Theorem 4.5

Assume that F_i *is convex for all* $i = 1, \ldots, m$. *Under Assumption 4.1, Algorithm 4.1 generates a sequence* $\{x^k\}$ *such that*

$$u_0(x^k) \leq \frac{\ell R}{2k} \quad \textit{for all } k \geq 1.$$

Proof. From Lemma 4.4 and the convexity of f_i and g_i, for all $x \in \mathbf{R}^n$ we have

$$\sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \leq \frac{\ell}{2} \left(\left\|x^k - x\right\|_2^2 - \left\|x^{k+1} - x\right\|_2^2\right).$$

Adding up the above inequality from $k = 0$ to $k = \hat{k}$, we obtain

$$\begin{aligned}\sum_{k=0}^{\hat{k}} \sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) &\leq \frac{\ell}{2} \left(\left\|x^0 - x\right\|_2^2 - \left\|x^{\hat{k}+1} - x\right\|_2^2\right) \\ &\leq \frac{\ell}{2} \left\|x^0 - x\right\|_2^2.\end{aligned}$$

Since $F_i(x^{\hat{k}+1}) \leq F_i(x^{k+1})$ for all $k \leq \hat{k}$ (see (4.7)), we get

$$\sum_{k=0}^{\hat{k}} \sum_{i=1}^m \lambda_i^k \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \leq \frac{\ell}{2} \left\|x^0 - x\right\|_2^2.$$

Let $\bar{\lambda}_i^{\hat{k}} := \sum_{k=0}^{\hat{k}} \lambda_i^k/(\hat{k}+1)$. Then, it follows that

$$\sum_{i=1}^{m} \bar{\lambda}_i^{\hat{k}} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \leq \frac{\ell}{2(\hat{k}+1)} \|x^0 - x\|_2^2.$$

Since $\bar{\lambda}_i^{\hat{k}} \geq 0$ and $\sum_{i=1}^{m} \bar{\lambda}_i^{\hat{k}} = 1$, we see that

$$\min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \leq \frac{\ell}{2(\hat{k}+1)} \|x^0 - x\|_2^2.$$

Therefore, we get

$$\sup_{F^* \in F(X^* \cap \Omega_F(F(x^0)))} \inf_{x \in F^{-1}(\{F^*\})} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \leq \frac{\ell R}{2(\hat{k}+1)}.$$

Thus, we obtain

$$\sup_{F^* \in F(X^* \cap \Omega_F(F(x^0)))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i^*\right) \leq \frac{\ell R}{2(\hat{k}+1)},$$

which gives

$$\sup_{x \in X^* \cap \Omega_F(F(x^0))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \leq \frac{\ell R}{2(\hat{k}+1)}. \tag{4.9}$$

Now, the inequality $F_i(x^k) \leq F_i(x^0)$ from (4.7) gives

$$\begin{aligned} &\sup_{x \in \Omega_F(F(x^0))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \\ &= \sup_{x \in \Omega_F(F(x^{\hat{k}+1}))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right), \end{aligned}$$

so we have

$$\begin{aligned} &\sup_{x \in \Omega_F(F(x^0))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right) \\ &\quad = \sup_{x \in \mathbf{R}^n} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x)\right). \end{aligned} \tag{4.10}$$

Moreover, from the assumption that for all $x \in \Omega_F(F(x^0))$ there exists $x^* \in X^*$

such that $F(x^*) \leq F(x)$, it follows that

$$\begin{aligned} &\sup_{x \in X^* \cap \Omega_F(F(x^0))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x) \right) \\ &\quad = \sup_{x \in \Omega_F(F(x^0))} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x) \right). \end{aligned} \tag{4.11}$$

Finally, from (4.9)–(4.11) we conclude that

$$u_0(x^{\hat{k}+1}) := \sup_{x \in \mathbf{R}^n} \min_{i=1,\ldots,m} \left(F_i(x^{\hat{k}+1}) - F_i(x) \right) \leq \frac{\ell R}{2(\hat{k}+1)}. \qquad \square$$

4.3.3 The strongly convex case

For (1.1) with $m = 1$, it is known that $\{x^k\}$ converges linearly to the optimal point when f_1 is strongly convex [**Beck2017**]. Now, we show that the same result holds for Algorithm 4.1.

Theorem 4.6
Let f_i and g_i have convexity parameters $\mu_i \in \mathbf{R}$ and $\nu_i \in \mathbf{R}$, respectively, and write $\mu := \min_{i=1,\ldots,m} \mu_i$ and $\nu := \min_{i=1,\ldots,m} \nu_i$. If $\ell > L$, then there exists a Pareto optimal point $x^* \in \mathbf{R}^n$ such that for each iteration k,

$$\left\| x^{k+1} - x^* \right\| \leq \sqrt{\frac{\ell - \mu}{\ell + \nu}} \left\| x^k - x^* \right\|.$$

Thus, we have

$$\left\| x^k - x^* \right\| \leq \left(\sqrt{\frac{\ell - \mu}{\ell + \nu}} \right)^k \left\| x^0 - x^* \right\|.$$

Proof. Since each F_i is strongly convex, the level set of every F_i is bounded. Thus, $\{x^k\}$ has an accumulation point $x^* \in \mathbf{R}^n$. Note that x^* is a Pareto optimal point [**Tanabe2019**]. Now, from Lemma 4.4, we have

$$\begin{aligned} \sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x^*) \right) \leq \frac{\ell}{2} &\left(\left\| x^k - x^* \right\|^2 - \left\| x^{k+1} - x^* \right\|^2 \right) \\ &- \frac{\mu}{2} \left\| x^k - x^* \right\|^2 - \frac{\nu}{2} \left\| x^{k+1} - x^* \right\|^2. \end{aligned}$$

Since the left-hand side is nonnegative because of (4.3), (4.6), and Lemma 4.2, we

obtain

$$0 \leq \frac{\ell}{2}\left(\left\|x^k - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2\right) - \frac{\mu}{2}\left\|x^k - x^*\right\|^2 - \frac{\nu}{2}\left\|x^{k+1} - x^*\right\|^2,$$

which is equivalent to

$$\left\|x^{k+1} - x^*\right\| \leq \sqrt{\frac{\ell - \mu}{\ell + \nu}}\left\|x^k - x^*\right\|. \qquad \square$$

4.3.4 The case that the multi-objective proximal-PL inequality is assumed

For (1.1) with $m = 1$ satisfying proximal-PL inequality, it is known that $\{F_1(x^k) - F_1^*\}$ converges linearly to zero with the proximal gradient method [**Karimi2016**]. Now, we show that $\{u_0(x^k)\}$ converges linearly to zero with Algorithm 4.1 for the multi-objective problem (1.1) that satisfies the multi-objective proximal-PL inequality (3.15).

Theorem 4.7
Suppose that (3.15) holds with a constant $\tau > 0$. Then, Algorithm 4.1 generates a sequence $\{x^k\}$ such that

$$u_0(x^{k+1}) \leq \left(1 - \frac{\tau}{\ell}\right) u_0(x^k).$$

Proof. Since ∇f_i is Lipschitz continuous with constant $\ell > L$, for all $i = 1, \ldots, m$ we get

$$\begin{aligned} F_i(x^{k+1}) - F_i(x^k) &\leq \langle \nabla f_i(x^k), d^k \rangle + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\left\|d^k\right\|^2 \\ &\leq \max_{i=1,\ldots,m}\left\{\langle \nabla f_i(x^k), d^k \rangle + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\left\|d^k\right\|^2\right\} \\ &= -w_\ell(x^k) \leq -\frac{\tau}{\ell} u_0(x^k), \end{aligned}$$

where the equality follows from (4.5), and the last inequality comes from (3.15). Then, for all $x \in \mathbf{R}^n$ we obtain

$$F_i(x^{k+1}) - F_i(x) \leq F_i(x^k) - F_i(x) - \frac{\tau}{\ell} u_0(x^k).$$

Therefore, it follows that

$$\sup_{x\in\mathbf{R}^n}\min_{i=1,\dots,m}\{F_i(x^{k+1})-F_i(x)\}\leq\sup_{x\in\mathbf{R}^n}\min_{i=1,\dots,m}\{F_i(x^k)-F_i(x)\}-\frac{\tau}{\ell}u_0(x^k),$$

which is equivalent to

$$u_0(x^{k+1})\leq\left(1-\frac{\tau}{\ell}\right)u_0(x^k).$$ □

Remark 4.4
While Theorem 4.6 implies linear convergence of $\{x^k\}$*, we show that* $\{u_0(x^k)\}$ *converges to zero linearly in Theorem 4.7. However, from Theorem 4.7, we can show that if each* f_i *is strongly convex with modulus* $\mu_i>0$*, then it follows that*

$$u_0(x^{k+1})\leq\left(1-\frac{L}{\ell\max\{L/\mu,1\}}\right)u_0(x^k),$$

with $\mu:=\max_{i=1,\dots,m}\mu_i$.

4.4 Application to robust multi-objective optimization

Now, let us apply the proposed algorithms to robust multi-objective optimization. Here, we suppose that the problems include uncertain parameters. Moreover, suppose that we can estimate the set of these uncertain parameters. Then, we try to optimize by considering the worst scenario. We observe that studies about robust multi-objective optimization is relatively new [**Ehrgott2014**, **Fliege2014**, **Morishita2016**].

Here, we consider the convex function g_i defined as follows:

$$g_i(x):=\max_{u\in\mathcal{U}_i}\hat{g}_i(x,u).\tag{4.12}$$

We call $\mathcal{U}_i\subseteq\mathbf{R}^n$ an *uncertainty set*. From now on, we assume $\mathcal{U}_i\subset\mathbf{R}^n$ and $\hat{g}_i\colon\mathbf{R}^n\times\mathbf{R}^n\to\mathbf{R}$ to be convex with respect to x. It is easy to see that g_i is also convex. However, g_i is not necessarily differentiable even if $\hat{g}_i$ is differentiable. First,

let us reformulate the subproblem (4.1) by using an extra variable $\gamma \in \mathbf{R}$ as

$$\begin{aligned} \min_{\gamma,d} \quad & \gamma + \frac{\ell}{2}\|d\|^2 \\ \text{s.t.} \quad & \nabla f_i(x)^\top d + g_i(x+d) - g_i(x) \leq \gamma, \quad i = 1, \ldots, m. \end{aligned}$$

Note that g_i is not easy to calculate, and thus, the subproblem is difficult to solve. When $\hat{g}_i$ and $\mathcal{U}_i$ have some special structure, the constraints can be written as explicit formulae by using the duality of (4.12). Now, assume that the dual problem of the maximization problem (4.12) is written as follows:

$$\begin{aligned} \min_{w_i} \quad & \tilde{g}_i(x, w_i) \\ \text{s.t.} \quad & w_i \in \tilde{\mathcal{U}}_i(x), \end{aligned}$$

where $\tilde{g}_i \colon \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ and $\tilde{\mathcal{U}}_i \colon \mathbf{R}^n \to 2^{\mathbf{R}^m}$. If strong duality holds, then we see that the subproblem (4.1) is equivalent to

$$\begin{aligned} \min_{\gamma,d,w_i} \quad & \gamma + \frac{\ell}{2}\|d\|^2 \\ \text{s.t.} \quad & \nabla f_i(x)^\top d + \tilde{g}_i(x+d, w_i) - g_i(x) \leq \gamma, \\ & w_i \in \tilde{\mathcal{U}}_i(x+d), \quad i = 1, \ldots, m. \end{aligned} \tag{4.13}$$

When $\tilde{g}_i$ and $\tilde{\mathcal{U}}_i$ have some explicit form, this problem is tractable. As we mention below, in this case, we can convert the above subproblem to some well-known convex optimization problems. This idea can be also seen in [**Ben-tal1998**]. In the following, we will introduce some robust multi-objective optimization problems where the subproblems can be written as quadratic programming, second-order cone programming or semidefinite programming problems.

4.4.1 Linearly constrained quadratic programming

Suppose that $\hat{g}_i(x, u) = u^\top x$ and $\mathcal{U}_i = \{u \in \mathbf{R}^n \mid A_i u \leq b_i\}$, where $A_i \in \mathbf{R}^{d \times n}$ and $b_i \in \mathbf{R}^d$, that is, $\hat{g}_i$ is linear in x, and $\mathcal{U}_i$ is a polyhedron. Suppose also that $\mathcal{U}_i$ is nonempty and bounded. Then, we can rewrite (4.12) as the following linear

programming problem:

$$\begin{array}{ll} \max\limits_{u} & x^\top u \\ \text{s.t.} & A_i u \le b_i. \end{array} \tag{4.14}$$

Its dual problem is given by

$$\begin{array}{ll} \min\limits_{w} & b_i^\top w \\ \text{s.t.} & A_i^\top w = x, \\ & w \ge 0. \end{array}$$

Since the strong duality holds, we can convert the subproblem (4.1) (or, equivalently (4.13)) to a linearly constrained quadratic programming problem:

$$\begin{array}{ll} \min\limits_{\gamma,d,w_i} & \gamma + \dfrac{\ell}{2}\|d\|^2 \\ \text{s.t.} & \nabla f_i(x)^\top d + b_i^\top w_i - g_i(x) \le \gamma, \\ & A_i^\top w_i = x + d, \\ & w_i \ge 0, \quad i = 1, \ldots, m. \end{array} \tag{4.15}$$

4.4.2 Second-order cone programming

Suppose that $\hat{g}_i(x,u) = u^\top x$ and $\mathcal{U}_i = \{a_i + P_i v \in \mathbf{R}^n \mid \|v\| \le 1,\ v \in \mathbf{R}^n\}$, where $a_i \in \mathbf{R}^n$ and $P_i \in \mathbf{R}^{n\times n}$, that is, $\hat{g}_i$ is once again linear in x and $\mathcal{U}_i$ is an ellipsoid. Then, for all $i = 1, \ldots, m$ we have

$$\begin{aligned} g_i(x) &= \max_{u\in\mathcal{U}_i} \hat{g}_i(x,u) \\ &= \max_{v:\|v\|\le 1} (a_i + P_i v)^\top x \\ &= a_i^\top x + \max_{v:\|v\|\le 1} (P_i^\top x)^\top v. \end{aligned}$$

If $P_i^\top x = 0$, then $\max\limits_{v:\|v\|\le 1} (P_i^\top x)^\top v = 0 = \|P_i^\top x\|$. If $P_i^\top x \ne 0$, then $\frac{P_i^\top x}{\|P_i^\top x\|}$ is a solution of $\max\limits_{v:\|v\|\le 1} (P_i^\top x)^\top v$, and hence $\max\limits_{v:\|v\|\le 1} (P_i^\top x)^\top v = \|P_i^\top x\|$. Consequently, we have

$$g_i(x) = a_i^\top x + \|P_i^\top x\|.$$

Therefore, introducing slack variables $\gamma \in \mathbf{R}$ and $\tau \in \mathbf{R}$, the subproblem (4.1) can be written as

$$\begin{aligned} \min_{\tau,\gamma,d} \quad & \tau \\ \text{s.t.} \quad & \nabla f_i(x)^\top d + \|P_i^\top (x+d)\| - \|P_i^\top x\| + a_i^\top d \le \gamma, \quad i = 1, \dots, m, \\ & \gamma + \frac{\ell}{2}\|d\|^2 \le \tau. \end{aligned}$$

Note that convex quadratic constraints can be converted to second-order cone constraints. Using the expression given in [**Alizadeh2003**], we get the following second-order cone programming problem (SOCP):

$$\begin{aligned} \min_{\tau,\gamma,d} \quad & \tau \\ \text{s.t.} \quad & \begin{bmatrix} -(\nabla f_i(x) + a_i)^\top d + \gamma + \|P_i^\top x\| \\ P_i^\top (x+d) \end{bmatrix} \in \mathcal{K}_{n+1}, \\ & \begin{bmatrix} 1 - \gamma + \tau \\ 1 + \gamma - \tau \\ \sqrt{2\ell} d \end{bmatrix} \in \mathcal{K}_{n+2}, \end{aligned} \tag{4.16}$$

where $\mathcal{K}_q := \{(y_0, \bar{y}) \in \mathbf{R} \times \mathbf{R}^{q-1} \mid y_0 \ge \|\bar{y}\|\}$ is the second-order cone in $\mathbf{R}^q$. The above SOCP can be solved efficiently with an interior point method [**Alizadeh2003**].

4.4.3 Semi-definite programming

Suppose that[2] $\hat{g}_i(x,u) = (x+u)^\top A_i (x+u)$ and $\mathcal{U}_i = \{a_i + P_i v \in \mathbf{R}^n \mid \|v\| \le 1\}$, where $A_i \in \mathbf{R}^{n\times n}$ and $A_i \succeq O$, $a_i \in \mathbf{R}^n$ and $P_i \in \mathbf{R}^{n\times n}$. Then, there exists a matrix $M_i \in \mathbf{R}^{n\times n}$ such that $A_i = M_i M_i^\top$. Note that $\hat{g}_i$ is convex quadratic and $\mathcal{U}_i$ is an ellipsoid. Here, without loss of generality we can assume that A is a symmetric matrix since $(x+u)^\top A_i (x+u) = (x+u)^\top \tilde{A}_i (x+u)$, where $\tilde{A}_i := (A_i + A_i^\top)/2$. Then, $g_i(x)$ can be given as

$$g_i(x) = \max_{v:\|v\|\le 1} (x + a_i + P_i v)^\top A_i (x + a_i + P_i v). \tag{4.17}$$

[2] We denote $A \succeq (\succ) O$ when A is positive semidefinite (positive definite). Also, $A \succeq (\succ) B$ if and only if $A - B \succeq (\succ) O$.

Since problem (4.17) is a maximization problem of a convex function, it is not a convex optimization problem. Fortunately, it can be seen as a subproblem of a trust region method, so its optimal value $g_i(x)$ can be obtained efficiently. Considering (4.17), we observe that

$$g_i(x+d) = \max_{v:\|v\|\leq 1} (x+d+a_i+P_iv)^\top A_i(x+d+a_i+P_iv). \tag{4.18}$$

From [**Beck2006**], the Lagrangian dual of the maximization problem (4.18) is given by

$$\begin{aligned} \min_{\alpha,w} \quad & -w \\ \text{s.t.} \quad & \begin{bmatrix} -P_i^\top A_i P_i & -P_i^\top A_i(x+d+a_i) \\ -(x+d+a_i)^\top A_i^\top P_i & -(x+d+a_i)^\top A_i(x+d+a_i) - w \end{bmatrix} \\ & \qquad\qquad\qquad\qquad \succeq \alpha \begin{bmatrix} -I_n & 0 \\ 0 & 1 \end{bmatrix}, \\ & \alpha \geq 0, \end{aligned} \tag{4.19}$$

where I_n stands for the identity matrix of dimension n. Let (α^*, w^*) be an optimal solution of (4.19) and assume that[3] $\dim(\ker(A_i + \alpha^* I_n)) \neq 1$. Since both (4.18) and (4.19) have strictly feasible solutions and $I_n \succ O$, then the strong duality holds from [**Beck2006**]. Therefore, recalling (4.13), the subproblem (4.1) is equivalent to

$$\begin{aligned} \min_{\gamma,d,w_i,\alpha_i} \quad & \gamma + \frac{\ell}{2}\|d\|^2 \\ \text{s.t.} \quad & \nabla f_i(x)^\top d - w_i - g_i(x) \leq \gamma, \\ & \begin{bmatrix} -P_i^\top A_i P_i + \alpha_i I_n & -P_i^\top A_i(x+d+a_i) \\ -(x+d+a_i)^\top A_i^\top P_i & -(x+d+a_i)^\top A_i(x+d+a_i) - w_i - \alpha_i \end{bmatrix} \\ & \qquad\qquad\qquad\qquad \succeq O, \\ & \alpha_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Now, by using slack variables $\tau \in \mathbf{R}$ and $\zeta_i \in \mathbf{R}$ and converting the convex quadratic constraints to second-order cone ones, we get the following semidefinite programming

[3] Here, dim denotes dimension of a space and ker means kernel of a matrix.

problem:

$$
\begin{aligned}
\min_{\tau,\alpha_i,w_i,\gamma,d} \quad & \tau \\
\text{s.t.} \quad & \nabla f_i(x)^\top d - w_i - g_i(x) \le \gamma, \\
& \begin{bmatrix} 1-\gamma+\tau \\ 1+\gamma-\tau \\ \sqrt{2\ell}d \end{bmatrix} \in \mathcal{K}_{n+2}, \\
& \begin{bmatrix} -P_i^\top A_i P_i + \alpha_i I_n & -P_i^\top A_i(x+d+a_i) \\ -(x+d+a_i)^\top A_i^\top P_i & \zeta_i \end{bmatrix} \succeq O, \\
& \begin{bmatrix} \dfrac{1-\zeta_i-w_i-\alpha_i}{2} \\ \dfrac{1+\zeta_i+w_i+\alpha_i}{2} \\ M_i^\top(x+d+a_i) \end{bmatrix} \in \mathcal{K}_{n+2}, \\
& \alpha_i \ge 0, \quad i=1,\dots,m,
\end{aligned}
\tag{4.20}
$$

where O stands for a zero matrix with appropriate dimension. Note that the second-order cone constraints can be converted further into semidefinite constraints.

4.5 Numerical experiments

In this section, we present some numerical results using Algorithm 4.1 for the problems in Section 4.4. The experiments are carried out on a machine with a 1.8GHz Intel Core i5 CPU and 8GB memory, and we implement all codes in MATLAB R2017a. We consider the problem (1.1), where $n=5$, $m=2$, $f_i(x)=\frac{1}{2}x^\top A_i x + a_i^\top x$, $g_i(x) = \max_{u\in\mathcal{U}_i} \hat{g}_i(x,u)$, $A_i \in \mathbf{R}^{n\times n}$, $a_i \in \mathbf{R}^n$, and $\hat{g}_i \colon \mathbf{R}^n \to \mathbf{R}$, $i=1,\dots,m$. Here, we assume that each A_i is positive semidefinite, so it can be decomposed as $A_i = M_i M_i^\top$, where $M_i \in \mathbf{R}^{n\times n}$. We generate M_i and a_i by choosing every component randomly from the standard normal distribution. To implement Algorithm 4.1, we make the following choices.

Remark 4.5

- *Every component of x^0 is chosen randomly from the standard normal distribution.*
- *In Experiments 1 and 3, we set the constant $\ell = 5$. In Experiment 2, we set the constant $\ell = 7$.*

- *The terminate criteria is replaced by* $\|d^k\| < \varepsilon := 10^{-6}$.

Also, we run each one of the following experiments 100 times from different initial points, and with $\delta = 0, 0.05, 0.1$. Naturally, when $\delta = 0$, no uncertainties are considered.

Experiment 1

In the first experiment, we solve the problem of Section 4.4.1. We assume that $g_i(x) = \max_{u \in \mathcal{U}_i} u^\top x$, $i = 1, 2$, where $\mathcal{U}_1 = \{u \in \mathbf{R}^5 \mid -\delta \leq u_i \leq \delta,\ i = 1, \ldots, 5\}$ and $\mathcal{U}_2 = \{u \in \mathbf{R}^5 \mid -\delta \leq (Bu)_i \leq \delta,\ i = 1, \ldots, 5\}$. Here, every component of $B \in \mathbf{R}^{5\times 5}$ is chosen randomly from the standard normal distribution and $\delta \geq 0$. We use the MATLAB solver *linprog* to solve (4.14) and *quadprog* to solve (4.15). Figure 4.1 is the result for this experiment. For each δ, we obtained part of the Pareto frontier, and as δ gets smaller the objective values become smaller.

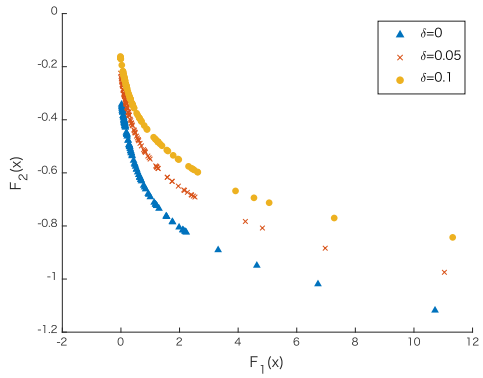


Figure 4.1: Result for Experiment 1

Experiment 2

In the second experiment, we solve the problem of Section 4.4.2. We assume that $g_i(x) = \max_{u \in \mathcal{U}_i} u^\top x$, where $\mathcal{U}_i = \{u \in \mathbf{R}^5 \mid \|u\| \leq \delta\}$, $i = 1, 2$. We use the MATLAB solver *SeDuMi* [**Sturm1999**] to solve (4.16). Figure 4.2 is the result for this experiment. Once again, we obtained part of the Pareto frontier for the problems with and without uncertainties.

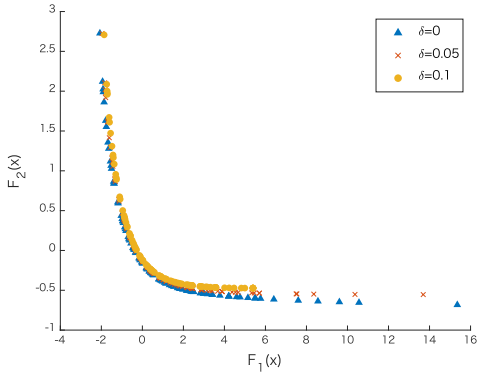


Figure 4.2: Result for Experiment 2

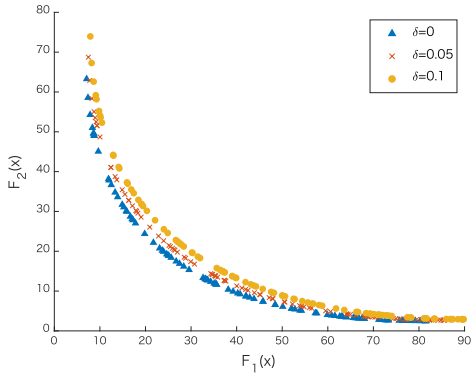


Figure 4.3: Result for Experiment 3

Experiment 3

Now, in the last experiment, we solve the problem of Section 4.4.3. We assume that $g_i(x) = \max_{u \in \mathcal{U}_i} (u+x)^\top BB^\top (u+x)$, where $\mathcal{U}_i = \{u \in \mathbf{R}^5 \mid \|u\| \leq \delta\}$, $i = 1, 2$. Here, once again, every component of $B \in \mathbf{R}^{5\times 5}$ is chosen randomly from the standard normal distribution and $\delta \geq 0$. We use the MATLAB solver *fmincon* to solve (4.17) and *SeDuMi* to solve (4.20). As it can be seen in Figure 4.3, we also obtained the Pareto frontier in this case.

4.6 Conclusions

We proposed the proximal gradient method for composite multi-objective optimization problems. Under reasonable assumptions, we proved the global convergence rate. Moreover, we presented some applications robust multi-objective optimization. In some robust optimization problems, we can convert the subproblems to well-known convex optimization problems. Finally, we carried out some numerical experiments for robust multi-objective optimization problems and we observed that the Pareto frontier changes when the uncertainty set is modified.

In recent years, faster methods such as Newton's method [**Fliege2009**] for differentiable multi-objective optimization problem have been also proposed. Therefore, an interesting topic for future research is to investigate the convergence rate of the proposed methods and to propose a proximal Newton-type algorithm for multi-objective optimization.

Chapter 5

An accelerated proximal gradient method for multi-objective optimization

5.1 Introduction

This chapter develops the accelerated proximal gradient method for the unconstrained comvex composite multi-objective optimization, i.e., (1.1) with (1.13), f_i is convex, and $C = \mathbf{R}^n$.

There are many studies related to the acceleration of single-objective first-order methods. After being established by Nesterov [**Nesterov1983**], researchers developed various accelerated schemes. In particular, the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [**Beck2009**], an accelerated version of the proximal gradient method, has contributed to a wide range of research fields, including image and signal processing. However, in the multi-objective case, the studies associated with accelerated algorithms are still insufficient. In 2020, El Moudden and El Mouatasim [**ElMoudden2020**] proposed an accelerated diagonal steepest descent method for multi-objective optimization, a natural extension of Nesterov's accelerated method for single-objective problems. They proved the global convergence rate of the algorithm ($O(1/k^2)$) under the assumption that the sequence of the Lagrange multipliers of the subproblems is eventually fixed. Nevertheless, this assumption is restrictive because it indicates that the approach is essentially the same as the (single-objective) Nesterov's method, only applied to the minimization of a weighted

sum of the objective functions.

Here, we propose a genuine accelerated proximal gradient method for multi-objective optimization. As it is usual, in each iteration, we solve a convex (scalar-valued) subproblem. While the accelerated and non-accelerated algorithms solve the same subproblem in the single-objective case, the subproblem of our accelerated method has terms that are not included in the non-accelerated version. However, we can ignore these terms in the single-objective case, and thus we can regard our proposed method as a generalization of FISTA. Moreover, under more natural assumptions, we prove the proposed method's global convergence rate ($O(1/k^2)$) by using a merit function (3.1) to measure the complexity.

Furthermore, having the practical computational efficiency in mind, we derive a dual of the subproblem, which is convex and differentiable. Such a dual problem turns out to be easier to solve than the original one, especially when the number of objective functions is smaller than the dimension of the decision variables. We can also reconstruct the original subproblem's solution directly from the dual optimum. In addition, we implement the whole algorithm using this dual problem and confirm its effectiveness with numerical experiments.

The outline of this paper is as follows. We present the accelerated proximal gradient method for multi-objective optimization in Section 5.2 and analyze its $O(1/k^2)$ convergence rate in Section 5.3. Moreover, Section 5.4 demonstrates the convergence of the iterates. Finally, we report some numerical results for test problems in Section 5.5, demonstrating that the proposed method is faster than the one without acceleration.

5.2 The algorithm

This section proposes an accelerated version of the proximal gradient method for multi-objective optimization. Similarly to the non-accelerated version given in the last section, a subproblem is considered in each iteration. More specifically, the proposed method solves the following subproblem for given $x \in \operatorname{dom} F$, $y \in \mathbf{R}^n$, and $\ell \geq L$:

$$\min_{z \in \mathbf{R}^n} \quad \varphi_\ell^{\mathrm{acc}}(z; x, y), \tag{5.1}$$

where

$$\varphi_\ell^{\mathrm{acc}}(z;x,y) \coloneqq \max_{i=1,\dots,m} [\langle \nabla f_i(y), z-y\rangle + g_i(z) + f_i(y) - F_i(x)] + \frac{\ell}{2}\|z-y\|^2.$$

Note that when $y = x$, (5.1) is reduced to the subproblem (4.1) of the proximal gradient method. Note also that when $m = 1$, the subproblem becomes

$$\min_{z\in\mathbf{R}^n} \quad \langle \nabla f_1(y), z-y\rangle + g_1(z) + \frac{\ell}{2}\|z-y\|^2, \tag{5.2}$$

which is the subproblem of the single-objective FISTA [**Beck2009**]. The distinctive feature of our proposal (5.1) is the term $f_i(y) - f_i(x)$, whereas the easy analogy from the single-objective subproblem (5.2) is

$$\min_{z\in\mathbf{R}^n} \quad \max_{i=1,\dots,m} [\langle \nabla f_i(y), z-y\rangle + g_i(z)] + \frac{\ell}{2}\|z-y\|^2.$$

By putting such a term, the inside of the max operator approximates $F_i(z) - F_i(x)$ rather than $F_i(z) - F_i(y)$. This is a negligible difference in the single-objective case, but deeply affects the proof in the multi-objective case.

Since g_i is convex for all $i = 1,\dots,m$, $z \mapsto \varphi_\ell^{\mathrm{acc}}(z;x,y)$ is strongly convex. Thus, the subproblem (5.1) has a unique optimal solution $p_\ell^{\mathrm{acc}}(x,y)$ and takes the optimal function value $\theta_\ell^{\mathrm{acc}}(x,y)$, i.e.,

$$p_\ell^{\mathrm{acc}}(x,y) \coloneqq \operatorname*{argmin}_{z\in\mathbf{R}^n} \varphi_\ell^{\mathrm{acc}}(z;x,y) \quad \text{and} \quad \theta_\ell^{\mathrm{acc}}(x,y) \coloneqq \min_{z\in\mathbf{R}^n} \varphi_\ell^{\mathrm{acc}}(z;x,y). \tag{5.3}$$

Moreover, the optimality condition of (5.1) implies that for all $x \in \operatorname{dom} F$ and $y \in \mathbf{R}^n$ there exists $\eta(x,y) \in \partial g(p_\ell^{\mathrm{acc}}(x,y))$ and a Lagrange multiplier $\lambda(x,y) \in \mathbf{R}^m$ such that

$$\sum_{i=1}^m \lambda_i(x,y)\,[\nabla f_i(y) + \eta_i(x,y)] = -\ell\,[p_\ell^{\mathrm{acc}}(x,y) - y] \tag{5.4a}$$

$$\lambda(x,y) \in \Delta^m, \quad \lambda_j(x,y) = 0 \quad \text{for all } j \notin \mathcal{I}(x,y), \tag{5.4b}$$

where Δ^m denotes the standard simplex (2.1) and

$$\mathcal{I}(x,y) \coloneqq \operatorname*{argmax}_{i=1,\dots,m} [\langle \nabla f_i(y), p_\ell^{\mathrm{acc}}(x,y) - y\rangle + g_i(p_\ell^{\mathrm{acc}}(x,y)) + f_i(y) - F_i(x)]\,. \tag{5.5}$$

Now, we introduce a relation useful for the subsequent analysis.

Lemma 5.1
Let p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$ be defined by (5.3). Then, we have

$$-\frac{\ell}{2}\left[\|p_\ell^{\mathrm{acc}}(x,y)-z\|^2-\|y-z\|^2\right] \geq \theta_\ell^{\mathrm{acc}}(x,y)+\sum_{i=1}^m \lambda_i(x,y)\left[\langle\nabla f_i(y),y-z\rangle-g_i(z)-f_i(y)+F_i(x)\right]$$

for all $x,z\in\operatorname{dom}F$ and $y\in\mathbf{R}^n$.

Proof. Let $x,z\in\operatorname{dom}F$ and $y\in\mathbf{R}^n$. From (5.4a) and the definition (2.5) of the subgradient, we get

$$\begin{aligned}
&-\ell\langle p_\ell^{\mathrm{acc}}(x,y)-y,p_\ell^{\mathrm{acc}}(x,y)-z\rangle\\
&\geq \sum_{i=1}^m \lambda_i(x,y)\left[\langle\nabla f_i(y),p_\ell^{\mathrm{acc}}(x,y)-z\rangle+g_i(p_\ell^{\mathrm{acc}}(x,y))-g_i(z)\right]\\
&= \sum_{i=1}^m \lambda_i(x,y)\left[\langle\nabla f_i(y),p_\ell^{\mathrm{acc}}(x,y)-y\rangle+g_i(p_\ell^{\mathrm{acc}}(x,y))+f_i(y)-F_i(x)\right]\\
&\quad+\sum_{i=1}^m \lambda_i(x,y)\left[\langle\nabla f_i(y),y-z\rangle-g_i(z)-f_i(y)+F_i(x)\right]\\
&= \max_{i=1,\ldots,m}\left[\langle\nabla f_i(y),p_\ell^{\mathrm{acc}}(x,y)-y\rangle+g_i(p_\ell^{\mathrm{acc}}(x,y))+f_i(y)-F_i(x)\right]\\
&\quad+\sum_{i=1}^m \lambda_i(x,y)\left[\langle\nabla f_i(y),y-z\rangle-g_i(z)-f_i(y)+F_i(x)\right],
\end{aligned}$$

where the second equality comes from (5.4b) and (5.5). Adding $(\ell/2)\|p_\ell^{\mathrm{acc}}(x,y)-y\|^2$ to both sides and the definition (5.3) of p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$ lead to

$$-\frac{\ell}{2}\left[2\langle p_\ell^{\mathrm{acc}}(x,y)-y,p_\ell^{\mathrm{acc}}(x,y)-z\rangle-\|p_\ell^{\mathrm{acc}}(x,y)-y\|^2\right] \geq \theta_\ell^{\mathrm{acc}}(x,y)+\sum_{i=1}^m \lambda_i(x,y)\left[\langle\nabla f_i(y),y-z\rangle-g_i(z)-f_i(y)+F_i(x)\right].$$

The left-hand side of this inequality is equal to

$$-\frac{\ell}{2}\left[2\langle p_\ell^{\mathrm{acc}}(x,y)-y,y-z\rangle+\|p_\ell^{\mathrm{acc}}(x,y)-y\|^2\right].$$

Hence, applying (5.11) with $(a, b, c) := (y, z, p_\ell^{\mathrm{acc}}(x, y))$, we get the desired inequality. □

We also note that by taking $z = y$ in the objective function of (5.1), we have

$$\theta_\ell^{\mathrm{acc}}(x, y) \leq \varphi_\ell^{\mathrm{acc}}(y; x, y) = \max_{i=1,\ldots,m} \{F_i(y) - F_i(x)\} \tag{5.6}$$

for all $x \in \operatorname{dom} F$ and $y \in \mathbf{R}^n$. Moreover, from Lemmas 2.2 and 2.3, and the fact that $\ell \geq L$, it follows that

$$\theta_\ell^{\mathrm{acc}}(x, y) \geq \max_{i=1,\ldots,m} \{F_i(p_\ell^{\mathrm{acc}}(x, y)) - F_i(x)\}$$

for all $x \in \operatorname{dom} F$ and $y \in \mathbf{R}^n$. We now characterize weak Pareto optimality in terms of the mappings p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$, similarly to Lemma 4.1 for the proximal gradient method.

Proposition 5.2
Let $p_\ell^{\mathrm{acc}}(x, y)$ and $\theta_\ell^{\mathrm{acc}}(x, y)$ be defined by (5.3). Then, the statements below hold.

(i) The following three conditions are equivalent:

- *(i) $y \in \mathbf{R}^n$ is weakly Pareto optimal for (1.1);*
- *(ii) $p_\ell^{\mathrm{acc}}(x, y) = y$ for some $x \in \mathbf{R}^n$;*
- *(iii) $\theta_\ell^{\mathrm{acc}}(x, y) = \max_{i=1,\ldots,m}[F_i(y) - F_i(x)]$ for some $x \in \mathbf{R}^n$.*

(ii) The mappings p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$ are locally Hölder continuous with exponent $1/2$ and locally Lipschitz continuous, respectively, i.e., for any bounded set $W \subseteq \mathbf{R}^n$, there exists $M_p > 0$ and $M_\theta > 0$ such that

$$\begin{aligned} \|p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - p_\ell^{\mathrm{acc}}(\check{x}, \check{y})\| &\leq M_p \|(\hat{x}, \hat{y}) - (\check{x}, \check{y})\|^{1/2}, \\ |\theta_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - \theta_\ell^{\mathrm{acc}}(\check{x}, \check{y})| &\leq M_\theta \|(\hat{x}, \hat{y}) - (\check{x}, \check{y})\| \end{aligned}$$

for all $\hat{x}, \hat{y}, \check{x}, \check{y} \in W$.

Proof. Claim (i) : From (5.6) and the fact that $\theta_\ell^{\mathrm{acc}}(x, y) = \varphi_\ell^{\mathrm{acc}}(p_\ell^{\mathrm{acc}}(x, y); x, y)$, the equivalence between (b) and (c) is apparent. Now, let us show that (a) and (b) are equivalent. When y is weakly Pareto optimal, we can immediately see from Lemma 4.1 that $p_\ell^{\mathrm{acc}}(x, y) = p_\ell(y) = y$ by letting $x = y$. Conversely, suppose

that $p_\ell^{\mathrm{acc}}(x,y) = y$ for some $x \in \mathbf{R}^n$. Let $z \in \mathbf{R}^n$ and $\alpha \in (0,1)$. The optimality of $p_\ell^{\mathrm{acc}}(x,y) = y$ for (5.1) gives

$$\begin{aligned}\max_{i=1,\dots,m}\{F_i(y) - F_i(x)\} &\le \varphi_\ell^{\mathrm{acc}}(y + \alpha(z-y); x, y)\\ &= \max_{i=1,\dots,m}\{\langle \nabla f_i(y), \alpha(z-y)\rangle + g_i(y+\alpha(z-y)) + f_i(y) - F_i(x)\}\\ &\qquad + \frac{\ell}{2}\|\alpha(z-y)\|^2.\end{aligned}$$

Thus, from the convexity of f_i, we get

$$\max_{i=1,\dots,m}\{F_i(y) - F_i(x)\} \le \max_{i=1,\dots,m}\{F_i(y+\alpha(z-y)) - F_i(x)\} + \frac{\ell}{2}\|\alpha(z-y)\|^2.$$

Moreover, the convexity of F_i yields

$$\begin{aligned}&\max_{i=1,\dots,m}\{F_i(y) - F_i(x)\}\\ &\qquad \le \max_{i=1,\dots,m}\{\alpha F_i(z) + (1-\alpha)F_i(y) - F_i(x)\} + \frac{\ell}{2}\|\alpha(z-y)\|^2\\ &\qquad \le \alpha \max_{i=1,\dots,m}\{F_i(z) - F_i(y)\} + \max_{i=1,\dots,m}\{F_i(y) - F_i(x)\} + \frac{\ell}{2}\|\alpha(z-y)\|^2.\end{aligned}$$

Therefore, we get

$$\max_{i=1,\dots,m}\{F_i(z) - F_i(y)\} \ge -\frac{\ell\alpha}{2}\|z-y\|^2.$$

Taking $\alpha \searrow 0$, we obtain

$$\max_{i=1,\dots,m}\{F_i(z) - F_i(y)\} \ge 0,$$

which implies the weak Pareto optimality of y.

Claim (ii) : Take $\hat{x}, \hat{y}, \check{x}, \check{y} \in W$. Adding the two inequalities of Lemma 5.1

with $(x, y, z) \coloneqq (\hat{x}, \hat{y}, p_\ell^{\mathrm{acc}}(\check{x}, \check{y})), (\check{x}, \check{y}, p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}))$ gives

$$
\begin{aligned}
&- \ell \|p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - p_\ell^{\mathrm{acc}}(\check{x}, \check{y})\|^2 + \frac{\ell}{2}\|p_\ell^{\mathrm{acc}}(\check{x}, \check{y}) - \hat{y}\|^2 + \frac{\ell}{2}\|p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - \check{y}\|^2 \\
&\quad \geq \theta_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) + \theta_\ell^{\mathrm{acc}}(\check{x}, \check{y}) \\
&\qquad + \sum_{i=1}^{m} \lambda_i(\hat{x}, \hat{y}) \left[\langle \nabla f_i(\hat{y}), \hat{y} - p_\ell^{\mathrm{acc}}(\check{x}, \check{y})\rangle - g_i(p_\ell^{\mathrm{acc}}(\check{x}, \check{y})) - f_i(\hat{y}) + F_i(\hat{x})\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(\check{x}, \check{y}) \left[\langle \nabla f_i(\check{y}), \check{y} - p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y})\rangle - g_i(p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y})) - f_i(\check{y}) + F_i(\check{x})\right].
\end{aligned}
$$

From the definition (5.3) of p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$ and (5.4b), we have

$$
\begin{aligned}
&- \ell \|p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - p_\ell^{\mathrm{acc}}(\check{x}, \check{y})\|^2 \\
&\quad \geq \sum_{i=1}^{m} \lambda_i(\check{x}, \check{y}) \left[\langle \nabla f_i(\hat{y}), p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - \hat{y}\rangle + g_i(p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y})) + f_i(\hat{y}) - F_i(\hat{x})\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(\hat{x}, \hat{y}) \left[\langle \nabla f_i(\check{y}), p_\ell^{\mathrm{acc}}(\check{x}, \check{y}) - \check{y}\rangle + g_i(p_\ell^{\mathrm{acc}}(\check{x}, \check{y})) + f_i(\check{y}) - F_i(\check{x})\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(\hat{x}, \hat{y}) \left[\langle \nabla f_i(\hat{y}), \hat{y} - p_\ell^{\mathrm{acc}}(\check{x}, \check{y})\rangle - g_i(p_\ell^{\mathrm{acc}}(\check{x}, \check{y})) - f_i(\hat{y}) + F_i(\hat{x})\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(\check{x}, \check{y}) \left[\langle \nabla f_i(\check{y}), \check{y} - p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y})\rangle - g_i(p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y})) - f_i(\check{y}) + F_i(\check{x})\right] \\
&\qquad - \frac{\ell}{2}\Big[\|p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - \check{y}\|^2 - \|p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - \hat{y}\|^2 \\
&\qquad\qquad + \|p_\ell^{\mathrm{acc}}(\check{x}, \check{y}) - \hat{y}\|^2 - \|p_\ell^{\mathrm{acc}}(\check{x}, \check{y}) - \check{y}\|^2\Big] \\
&\quad = \sum_{i=1}^{m} \lambda_i(\hat{x}, \hat{y})[\langle \nabla f_i(\hat{y}), \hat{y} - \check{y}\rangle + \langle \nabla f_i(\hat{y}) - \nabla f_i(\check{y}), \check{y} - p_\ell^{\mathrm{acc}}(\check{x}, \check{y})\rangle \\
&\qquad\qquad - f_i(\hat{y}) + f_i(\check{y}) + F_i(\hat{x}) - F_i(\check{x})] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(\check{x}, \check{y})[\langle \nabla f_i(\check{y}), \check{y} - \hat{y}\rangle + \langle \nabla f_i(\check{y}) - \nabla f_i(\hat{y}), \hat{y} - p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y})\rangle \\
&\qquad\qquad - f_i(\check{y}) + f_i(\hat{y}) + F_i(\check{x}) - F_i(\hat{x})] \\
&\qquad - \ell\langle p_\ell^{\mathrm{acc}}(\hat{x}, \hat{y}) - p_\ell^{\mathrm{acc}}(\check{x}, \check{y}), \hat{y} - \check{y}\rangle.
\end{aligned}
$$

Thus, (5.4b) and Cauchy-Schwarz inequalities applied in each inner product that

appears in the right-hand side of the above expression imply

$$\begin{aligned}
&-\ell\|p_\ell^{\mathrm{acc}}(\hat{x},\hat{y})-p_\ell^{\mathrm{acc}}(\check{x},\check{y})\|^2\\
&\geq -2\max_{i=1,\dots,m}\|\nabla f_i(\hat{y})\|\|\hat{y}-\check{y}\|\\
&\quad-\left[\|\hat{y}-p_\ell^{\mathrm{acc}}(\hat{x},\hat{y})\|+\|\check{y}-p_\ell^{\mathrm{acc}}(\check{x},\check{y})\|\right]\max_{i=1,\dots,m}\|\nabla f_i(\hat{y})-\nabla f_i(\check{y})\|\\
&\quad-2\max_{i=1,\dots,m}|f_i(\hat{y})-f_i(\check{y})|-2\max_{i=1,\dots,m}|F_i(\hat{x})-F_i(\check{x})|\\
&\quad-\ell\|p_\ell^{\mathrm{acc}}(\hat{x},\hat{y})-p_\ell^{\mathrm{acc}}(\check{x},\check{y})\|\|\hat{y}-\check{y}\|.
\end{aligned}$$

Let us now show that each term of the right-hand side of the above inequality is bounded by a positive constant multiple of $-\|\hat{x}-\check{x}\|$ or $-\|\hat{y}-\check{y}\|$. The first term is direct because the boundedness of W implies $\max_{i=1,\dots,m}\|\nabla f_i(\hat{y})\|<+\infty$. Since W is bounded and the objective function of (5.1) is strongly convex, $p_\ell^{\mathrm{acc}}(x,y)$ also belongs to some bounded set for all $x,y\in W$, thus $\|\hat{y}-p_\ell^{\mathrm{acc}}(\hat{x},\hat{y})\|<+\infty$ and $\|\check{y}-p_\ell^{\mathrm{acc}}(\check{x},\check{y})\|<+\infty$. Thus, the Lipschitz continuity of ∇f_i shows such a boundedness of the second term. Moreover, the locally Lipschitz continuity of f_i and F_i derived by the continuous differentiability of f_i and convexity F_i lead to the similar property for the third and fourth terms. Hence, p_ℓ^{acc} is Hölder continuous with exponent $1/2$ on W.

On the other hand, the definition (5.3) of p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$ gives

$$\begin{aligned}
\theta_\ell^{\mathrm{acc}}(\hat{x},\hat{y}) - \theta_\ell^{\mathrm{acc}}(\check{x},\check{y}) &\le \varphi_\ell^{\mathrm{acc}}(p_\ell^{\mathrm{acc}}(\check{x},\check{y});\hat{x},\hat{y}) - \varphi_\ell^{\mathrm{acc}}(p_\ell^{\mathrm{acc}}(\check{x},\check{y});\check{x},\check{y}) \\
&= \max_{i=1,\dots,m} [\langle \nabla f_i(\hat{y}), p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\rangle + g_i(p_\ell^{\mathrm{acc}}(\check{x},\check{y})) + f_i(\hat{y}) - F_i(\hat{x})] \\
&\quad - \max_{i=1,\dots,m} [\langle \nabla f_i(\check{y}), p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \check{y}\rangle + g_i(p_\ell^{\mathrm{acc}}(\check{x},\check{y})) + f_i(\check{y}) - F_i(\check{x})] \\
&\quad + \frac{\ell}{2}\Big[\|p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\|^2 - \|p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \check{y}\|^2\Big] \\
&\le \max_{i=1,\dots,m} [\langle \nabla f_i(\check{y}), \check{y} - \hat{y}\rangle + \langle \nabla f_i(\hat{y}) - \nabla f_i(\check{y}), p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\rangle \\
&\qquad\qquad + f_i(\hat{y}) - f_i(\check{y}) - F_i(\hat{x}) + F_i(\check{x})] \\
&\quad + \frac{\ell}{2}\langle 2p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y} - \check{y}, \check{y} - \hat{y}\rangle \\
&\le \max_{i=1,\dots,m} \|\nabla f_i(\check{y})\|\|\hat{y} - \check{y}\| + \|\hat{y} - p_\ell^{\mathrm{acc}}(\check{x},\check{y})\| \max_{i=1,\dots,m} \|\nabla f_i(\hat{y}) - \nabla f_i(\check{y})\| \\
&\quad + \max_{i=1,\dots,m} |f_i(\hat{y}) - f_i(\check{y})| + \max_{i=1,\dots,m} |F_i(\hat{x}) - F_i(\check{x})| \\
&\quad + \frac{\ell}{2}\|2p_\ell^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y} - \check{y}\|\|\hat{y} - \check{y}\|,
\end{aligned}$$

where the second inequality follows from the relation

$$\max_{i=1,\dots,m} a_i - \max_{i=1,\dots,m} b_i \le \max_{i=1,\dots,m} (a_i - b_i) \quad \text{for all } a, b \in \mathbf{R}^m,$$

and the third inequality comes from (5.4b) and Cauchy-Schwarz inequalites. Since the above inequality holds even if we interchange $(\hat{x},\hat{y})$ and $(\check{x},\check{y})$, we can show the Lipschitz continuity of $\theta_\ell^{\mathrm{acc}}$ on W in the same way as in the previous paragraph. □

Note that the Hölder exponent $1/2$ mentioned in Proposition 5.2 (ii) is optimal, i.e., for some F_i, p_ℓ^{acc} is not Hölder continuous with exponent $\alpha > 1/2$. In fact, this result was also proved for multi-objective steepest direction in [**Svaiter2018**].

Proposition 5.2 suggests that we can use $\|p_\ell^{\mathrm{acc}}(x,y) - y\|_\infty < \varepsilon$ for some $\varepsilon > 0$ as a stopping criteria. Now, we state below the proposed algorithm.

The sequence $\{t_k\}$ defined in lines 2 and 5 of Algorithm 5.1 generalizes the well-known momentum factors in single-objective accelerated methods. For example, when $a = 0$ and $b = 1/4$, they coincide with the one in Algorithm 5.1 and the original FISTA [**Nesterov1983**, **Beck2009**] ($t_1 = 1$ and $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$). Moreover, if $b = a^2/4$, then $\{t_k\}$ has the general term $t_k = (1-a)k/2 + (1+a)/2$, which corresponds to the one used in [**Chambolle2015**, **Su2016**, **Attouch2016**,

Algorithm 5.1 Accelerated proximal gradient method with general stepsizes for (1.1)

Input: Set $x^0 = y^1 \in \operatorname{dom} F, \ell \geq L, \varepsilon > 0, a \in [0,1), b \in [a^2/4, 1/4]$.
Output: x^*: A weakly Pareto optimal point

1: $k \leftarrow 1$
2: $t_1 \leftarrow 1$
3: **while** $\left\|p_\ell^{\text{acc}}(x^{k-1}, y^k) - y^k\right\|_\infty \geq \varepsilon$ **do**
4: $\quad x^k \leftarrow p_\ell^{\text{acc}}(x^{k-1}, y^k)$
5: $\quad t_{k+1} \leftarrow \sqrt{t_k^2 - at_k + b} + 1/2$
6: $\quad \gamma_k \leftarrow (t_k - 1)/t_{k+1}$
7: $\quad y^{k+1} \leftarrow x^k + \gamma_k(x^k - x^{k-1})$
8: $\quad k \leftarrow k + 1$
9: **end while**

Attouch2018]. This means that our generalization allows a finer tuning of the algorithm by varying a and b. We show below some properties of $\{t_k\}$ and $\{\gamma_k\}$.

Lemma 5.3
Let $\{t_k\}$ and $\{\gamma_k\}$ be defined by lines 2, 5 and 6 in Algorithm 5.1 for arbitrary $a \in [0,1)$ and $b \in [a^2/4, 1/4]$. Then, the following inequalities hold for all $k \geq 1$.

(i) $t_{k+1} \geq t_k + \dfrac{1-a}{2}$ *and* $t_k \geq \dfrac{1-a}{2}k + \dfrac{1+a}{2}$;

(ii) $t_{k+1} \leq t_k + \dfrac{1-a+\sqrt{4b-a^2}}{2}$ *and* $t_k \leq \dfrac{1-a+\sqrt{4b-a^2}}{2}(k-1) + 1 \leq k$;

(iii) $t_k^2 - t_{k+1}^2 + t_{k+1} = at_k - b + \dfrac{1}{4} \geq at_k$;

(iv) $0 \leq \gamma_k \leq \dfrac{k-1}{k+1/2}$;

(v) $1 - \gamma_k^2 \geq \dfrac{1}{t_k}$.

Proof. Claim (i) : From the definition of $\{t_k\}$, we have

$$t_{k+1} = \sqrt{t_k^2 - at_k + b} + \frac{1}{2} = \sqrt{\left(t_k - \frac{a}{2}\right)^2 + \left(b - \frac{a^2}{4}\right)} + \frac{1}{2}. \tag{5.7}$$

Since $b \geq a^2/4$, we get

$$t_{k+1} \geq \left|t_k - \frac{a}{2}\right| + \frac{1}{2}.$$

Since $t_1 = 1 \geq a/2$, we can quickly see that $t_k \geq a/2$ for any k by induction. Thus, we have

$$t_{k+1} \geq t_k + \frac{1-a}{2}.$$

Applying the above inequality recursively, we obtain

$$t_k \geq \frac{1-a}{2}(k-1) + t_1 = \frac{1-a}{2}k + \frac{1+a}{2}.$$

Claim (ii) : From (5.7) and the relation $\sqrt{\alpha+\beta} \leq \sqrt{\alpha} + \sqrt{\beta}$ with $\alpha, \beta \geq 0$, we get the first inequality. Using it recursively, it follows that

$$t_k \leq \frac{1-a+\sqrt{4b-a^2}}{2}(k-1) + t_1 = \frac{1-a+\sqrt{4b-a^2}}{2}(k-1) + 1.$$

Since $a \in [0,1), b \in [a^2/4, 1/4]$, we observe that

$$\frac{1-a+\sqrt{4b-a^2}}{2} \leq \frac{1-a+\sqrt{1-a^2}}{2} \leq 1.$$

Hence, the above two inequalities lead to the desired result.

Claim (iii) : An easy computation shows that

$$\begin{aligned} t_k^2 - t_{k+1}^2 + t_{k+1} &= t_k^2 - \left[\sqrt{t_k^2 - at_k + b} + \frac{1}{2}\right]^2 + \sqrt{t_k^2 - at_k + b} + \frac{1}{2} \\ &= at_k - b + \frac{1}{4} \geq at_k, \end{aligned}$$

where the inequality holds since $b \leq 1/4$.

Claim (iv) : The first inequlity is clear from the definition of γ_k since claim (i) yields $t_k \geq 1$. Again, the definition of γ_k and claim (i) give

$$\gamma_k = \frac{t_k - 1}{t_{k+1}} \leq \frac{t_k - 1}{t_k + (1-a)/2} = 1 - \frac{3-a}{2t_k + 1 - a}.$$

Combining with claim (ii) , we get

$$\begin{aligned}\gamma_k &\leq 1 - \frac{3-a}{\left(1-a+\sqrt{4b-a^2}\right)(k-1)+3-a} \\ &= \frac{\left(1-a+\sqrt{4b-a^2}\right)(k-1)}{\left(1-a+\sqrt{4b-a^2}\right)(k-1)+3-a} \\ &= \frac{k-1}{k-1+(3-a)/\left(1-a+\sqrt{4b-a^2}\right)}.\end{aligned} \tag{5.8}$$

On the other hand, it follows that

$$\min_{a\in[0,1),b\in[a^2/4,1/4]} \frac{3-a}{1-a+\sqrt{4b-a^2}} = \min_{a\in[0,1)} \frac{3-a}{1-a+\sqrt{1-a^2}} = \frac{3}{2}, \tag{5.9}$$

where the second equality follows from the monotonic non-decreasing property implied by

$$\frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{3-a}{1-a+\sqrt{1-a^2}}\right) = \frac{2\sqrt{1-a^2}+3a-1}{\left(\sqrt{1-a^2}-a+1\right)^2\sqrt{1-a^2}} > 0 \quad \text{for all } a\in[0,1).$$

Combining (5.8) and (5.9), we obtain $\gamma_k \leq (k-1)/(k+1/2)$.

Claim (v) : Claim (i) implies that $t_{k+1} > t_k \geq 1$. Thus, the definition of γ_k implies that

$$1-\gamma_k^2 = 1-\left(\frac{t_k-1}{t_{k+1}}\right)^2 \geq 1-\left(\frac{t_k-1}{t_k}\right)^2 = \frac{2t_k-1}{t_k^2} \geq \frac{2t_k-t_k}{t_k^2} = \frac{1}{t_k}. \qquad \square$$

We end this section by noting some remarks about the proposed algorithm.

Remark 5.1

(i) Since $x \in \operatorname{dom} F$ implies $p_\ell^{\mathrm{acc}}(x,y) \in \operatorname{dom} F$, every x^k computed by the above algorithm is in $\operatorname{dom} F$. However, y^k is not necessarily in $\operatorname{dom} F$.

(ii) Since $y^1 = x^0$, it follows from (5.6) that

$$\theta_\ell^{\mathrm{acc}}(x^0, y^1) \leq 0,$$

but the inequality $\theta_\ell^{\mathrm{acc}}(x^{k-1}, y^k) \leq 0$ does not necessarily hold for $k \geq 2$.

(iii) When $m = 1$, we can remove the term $f_i(y) - F_i(x)$ from the subproblem (5.1), so Algorithm 5.1 corresponds to the Fast Iterative Shrinkage-Thresholding Al-

*gorithm (FISTA) [***Beck2009***] for single-objective optimization.*

*(iv) Algorithm 5.1 induces the accelerated versions of first-order algorithms such as the steepest descent [***Fliege2000***], proximal point [***Bonnel2005***], and projected gradient methods [***Grana-Drummond2004***].*

(v) Even if it is difficult to estimate L, we can update the constant ℓ to satisfy $F_i(p_\ell^{\mathrm{acc}}(x^{k-1}, y^k)) - F_i(x^{k-1}) \leq \theta_\ell^{\mathrm{acc}}(x^{k-1}, y^k)$ for all $i = 1, \ldots, m$ in each iteration by a finite number of backtracking steps. Moreover, we can restrict the assumption of ∇f_i's Lipschitz continuity on the level set $\mathbf{lev}_F(F(x^0))$ without affecting the analysis in the subsequent sections.

5.3 Convergence rates analysis

This section shows that Algorithm 5.1 has the $O(1/k^2)$ convergence rate also holds in that case. We present below the main theorem of this section.

Theorem 5.4

Let $\{x^k\}$ be a sequence generated by Algorithm 5.1 and recall that u_0 is given by (3.1). Then, the following two equations hold:

(i) $F_i(x^k) \leq F_i(x^0)$ for all $i = 1, \ldots, m$ and $k \geq 0$;

(ii) $u_0(x^k) = O(1/k^2)$ as $k \to \infty$ under Assumption 4.1.

Claim (i) means that $\{x^k\} \subseteq \mathbf{lev}_F(F(x^0))$, where $\mathbf{lev}_F$ denotes the level set of F (cf. (2.2)). Note, however, that the objective functions are generally not monotonically non-increasing. Claim (ii) also claims the global convergence rate.

Before proving Theorem 5.4, let us give several lemmas. First, we present some properties of $\{t_k\}$ and $\{\gamma_k\}$. As in [**Tanabe2022a**], we also introduce $\sigma_k \colon \mathbf{R}^n \to \mathbf{R} \cup \{-\infty\}$ and $\rho_k \colon \mathbf{R}^n \to \mathbf{R}$ for $k \geq 0$ as follows, which assist the analysis:

$$\begin{aligned} \sigma_k(z) &:= \min_{i=1,\ldots,m} \left[F_i(x^k) - F_i(z)\right], \\ \rho_k(z) &:= \left\| t_{k+1} x^{k+1} - (t_{k+1} - 1) x^k - z \right\|^2. \end{aligned} \tag{5.10}$$

The following lemma on σ_k is helpful in the subsequent discussions.

Lemma 5.5

*[***Tanabe2022a***] Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 5.1. Then, the following inequalities hold for all $z \in \mathbf{R}^n$ and $k \geq 0$:*

$$(i)\ \sigma_{k+1}(z) \leq -\frac{\ell}{2}\left(2\langle x^{k+1}-y^{k+1}, y^{k+1}-z\rangle + \left\|x^{k+1}-y^{k+1}\right\|^2\right) - \frac{\ell-L}{2}\left\|x^{k+1}-y^{k+1}\right\|^2;$$

$$(ii)\ \sigma_k(z)-\sigma_{k+1}(z) \geq \frac{\ell}{2}\left(2\langle x^{k+1}-y^{k+1}, y^{k+1}-x^k\rangle + \left\|x^{k+1}-y^{k+1}\right\|^2\right) + \frac{\ell-L}{2}\left\|x^{k+1}-y^{k+1}\right\|^2.$$

Therefore, from Lemma 5.3 (v), we can obtain the following result quickly in the same way as in the proof of [**Tanabe2022a**].

Lemma 5.6
Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 5.1. Then, we have

$$\sigma_{k_1}(z)-\sigma_{k_2}(z) \geq \frac{\ell}{2}\left(\left\|x^{k_2}-x^{k_2-1}\right\|^2 - \left\|x^{k_1}-x^{k_1-1}\right\|^2 + \sum_{k=k_1}^{k_2-1}\frac{1}{t_k}\left\|x^k-x^{k-1}\right\|^2\right)$$

for any $k_2 \geq k_1 \geq 1$.

We can now show the first part of Theorem 5.4.

Proof of Theorem 5.4 (i). From Lemma 5.6, we can prove this part with similar arguments used in the proof of [**Tanabe2022a**]. □

The next step is to prepare the proof of Theorem 5.4 (ii). First, we mention the following relation, used frequently hereafter:

$$\left\|v^2-v^1\right\|^2 + 2\langle v^2-v^1, v^1-v^3\rangle = \left\|v^2-v^3\right\|^2 - \left\|v^1-v^3\right\|^2, \tag{5.11}$$

$$\sum_{s=1}^{r}\sum_{p=1}^{s} A_p = \sum_{p=1}^{r}\sum_{s=p}^{r} A_p \tag{5.12}$$

for any vectors v^1, v^2, v^3 and sequence $\{A_p\}$. With these, we show the lemma below, which is similar to [**Tanabe2022a**] but more complex due to the generalization of $\{t_k\}$.

Lemma 5.7
Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 5.1. Also, let σ_k and ρ_k be

defined by (5.10). Then, we have

$$\begin{aligned}
&\frac{\ell}{2}\left\|x^0 - z\right\|^2 \\
&\quad \geq \frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4} - b\right)k\right]\sigma_{k+1}(z) \\
&\qquad + \frac{\ell}{2(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left\|x^{k+1} - x^k\right\|^2 \\
&\qquad + \frac{\ell}{2(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|^2 \\
&\qquad + \frac{\ell}{2}\rho_k(z) + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2\left\|x^{p+1} - y^{p+1}\right\|^2
\end{aligned}$$

for all $k \geq 0$ and $z \in \mathbf{R}^n$.

Proof. Let $p \geq 1$ and $z \in \mathbf{R}^n$. Recall that Lemma 5.5 gives

$$\begin{aligned}
-\sigma_{p+1}(z) &\geq \frac{\ell}{2}\left[2\left\langle x^{p+1} - y^{p+1}, y^{p+1} - z\right\rangle + \left\|x^{p+1} - y^{p+1}\right\|^2\right] \\
&\qquad + \frac{\ell - L}{2}\left\|x^{p+1} - y^{p+1}\right\|^2, \\
\sigma_p(z) - \sigma_{p+1}(z) &\geq \frac{\ell}{2}\left[2\left\langle x^{p+1} - y^{p+1}, y^{p+1} - x^p\right\rangle + \left\|x^{p+1} - y^{p+1}\right\|^2\right] \\
&\qquad + \frac{\ell - L}{2}\left\|x^{p+1} - y^{p+1}\right\|^2.
\end{aligned}$$

We then multiply the second inequality above by $(t_{p+1} - 1)$ and add it to the first one:

$$\begin{aligned}
&(t_{p+1} - 1)\sigma_p(z) - t_{p+1}\sigma_{p+1}(z) \\
&\quad \geq \frac{\ell}{2}\left[t_{p+1}\left\|x^{p+1} - y^{p+1}\right\|^2 + 2\left\langle x^{p+1} - y^{p+1}, t_{p+1}y^{p+1} - (t_{p+1} - 1)x^p - z\right\rangle\right] \\
&\qquad + \frac{\ell - L}{2}t_{p+1}\left\|x^{p+1} - y^{p+1}\right\|^2.
\end{aligned}$$

Multiplying this inequality by t_{p+1} and using the relation $t_p^2 = t_{p+1}^2 - t_{p+1} + (at_p -$

$b + 1/4$) (cf. Lemma 5.3 (iii)), we get

$$t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \geq \frac{\ell}{2}\Big[\big\|t_{p+1}(x^{p+1} - y^{p+1})\big\|^2 + 2t_{p+1}\big\langle x^{p+1} - y^{p+1}, t_{p+1}y^{p+1} - (t_{p+1}-1)x^p - z\big\rangle\Big] + \frac{\ell - L}{2}t_{p+1}^2\big\|x^{p+1} - y^{p+1}\big\|^2 + \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).$$

Applying (5.11) to the right-hand side of the last inequality with

$$v^1 := t_{p+1}y^{p+1}, \quad v^2 := t_{p+1}x^{p+1}, \quad v^3 := (t_{p+1}-1)x^p + z.$$

we get

$$t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \geq \frac{\ell}{2}\left[\big\|t_{p+1}x^{p+1} - (t_{p+1}-1)x^p - z\big\|^2 - \big\|t_{p+1}y^{p+1} - (t_{p+1}-1)x^p - z\big\|^2\right] + \frac{\ell - L}{2}t_{p+1}^2\big\|x^{p+1} - y^{p+1}\big\|^2 + \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).$$

Recall that $\rho_p(z) := \|t_{p+1}x^{p+1} - (t_{p+1}-1)x^p - z\|^2$. Then, considering the definition of y^p given in line 7 of Algorithm 5.1, we obtain

$$t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \geq \frac{\ell}{2}\left[\rho_p(z) - \rho_{p-1}(z)\right] + \frac{\ell - L}{2}t_{p+1}^2\big\|x^{p+1} - y^{p+1}\big\|^2 + \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).$$

Now, let $k \geq 0$. Lemma 5.6 with $(k_1, k_2) = (p, k+1)$ implies

$$t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \geq \frac{\ell}{2}\left[\rho_p(z) - \rho_{p-1}(z)\right] + \frac{\ell - L}{2}t_{p+1}^2\big\|x^{p+1} - y^{p+1}\big\|^2 + \left(at_p - b + \frac{1}{4}\right)\left[\sigma_{k+1}(z) + \frac{\ell}{2}\left(\big\|x^{k+1} - x^k\big\|^2 - \big\|x^p - x^{p-1}\big\|^2 + \sum_{r=p}^{k}\frac{1}{t_r}\big\|x^r - x^{r-1}\big\|^2\right)\right].$$

Adding up the above inequality from $p = 1$ to $p = k$, the fact that $t_1 = 1$ and $\rho_0(z) =$

$\|x^1 - z\|^2$ leads to

$$\begin{aligned}
&\sigma_1(z) - t_{k+1}^2 \sigma_{k+1}(z) \\
&\quad \geq \frac{\ell}{2}\left[\rho_k(z) - \left\|x^1 - z\right\|^2\right] + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{k+1}^2 \left\|x^{k+1} - y^{k+1}\right\|^2 \\
&\qquad + \left(a\sum_{p=1}^{k} t_p + \left(\frac{1}{4} - b\right)k\right)\left[\sigma_{k+1}(z) + \frac{\ell}{2}\left\|x^{k+1} - x^k\right\|^2\right] \\
&\qquad\quad - \frac{\ell}{2}\sum_{p=1}^{k}\left(at_p - b + \frac{1}{4}\right)\left\|x^p - x^{p+1}\right\|^2 \\
&\qquad\quad + \frac{\ell}{2}\sum_{p=1}^{k}\left(at_p - b + \frac{1}{4}\right)\sum_{r=p}^{k}\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|^2. \quad (5.13)
\end{aligned}$$

Let us write the last two terms of the right-hand side for (5.13) as S_1 and S_2, respectively. Eq. (5.12) yields

$$\begin{aligned}
S_2 &= \frac{\ell}{2}\sum_{r=1}^{k}\sum_{p=1}^{r}\left(at_p - b + \frac{1}{4}\right)\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|^2 \\
&= \frac{\ell}{2}\sum_{p=1}^{k}\sum_{r=1}^{p}\left(at_r - b + \frac{1}{4}\right)\frac{1}{t_p}\left\|x^p - x^{p-1}\right\|^2.
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
S_1 + S_2 &= \frac{\ell}{2}\sum_{p=1}^{k}\left[\frac{1}{t_p}\sum_{r=1}^{p}\left(at_r - b + \frac{1}{4}\right) - \left(at_p - b + \frac{1}{4}\right)\right]\left\|x^p - x^{p-1}\right\|^2 \\
&= \frac{\ell}{2}\sum_{p=1}^{k}\frac{1}{t_p}\left[a\left(\sum_{r=1}^{p-1} t_r - t_p^2 + t_p\right) + \left(\frac{1}{4} - b\right)(p - t_p)\right]\left\|x^p - x^{p-1}\right\|^2. \quad (5.14)
\end{aligned}$$

Again $t_1 = 1$ gives

$$\begin{aligned}
-t_p^2 + t_p &= \sum_{r=1}^{p-1}(-t_{r+1}^2 + t_{r+1} + t_r^2 - t_r) = \sum_{r=1}^{p-1}\left(-(1-a)t_r - b + \frac{1}{4}\right) \\
&= -(1-a)\sum_{r=1}^{p-1} t_r + \left(\frac{1}{4} - b\right)(p-1),
\end{aligned}$$

where the second equality comes from Lemma 5.3 (iii). Thus, we get

$$\sum_{r=1}^{p-1} t_r = \frac{t_p^2 - t_p}{1-a} + \left(\frac{1}{4} - b\right)\frac{p-1}{1-a}. \tag{5.15}$$

Substituting this into (5.14), it follows that

$$\begin{aligned} &S_1 + S_2 \\ &\quad = \frac{\ell}{2(1-a)} \sum_{p=1}^{k} \left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right] \left\|x^p - x^{p-1}\right\|^2. \end{aligned}$$

Combined with (5.13) and (5.15), we have

$$\begin{aligned} &\sigma_1(z) - t_{k+1}^2 \sigma_{k+1}(z) \\ &\quad \geq \frac{\ell}{2}\left[\rho_k(z) - \left\|x^1 - z\right\|^2\right] + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2 \left\|x^{k+1} - y^{k+1}\right\|^2 \\ &\qquad + \frac{1}{1-a}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left[\sigma_{k+1}(z) + \frac{\ell}{2}\left\|x^{k+1} - x^k\right\|^2\right] \\ &\qquad + \frac{\ell}{2(1-a)} \sum_{p=1}^{k} \left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right] \left\|x^p - x^{p-1}\right\|^2. \end{aligned}$$

Easy calculations give

$$\begin{aligned} &\sigma_1(z) + \frac{\ell}{2}\left\|x^1 - z\right\|^2 \\ &\quad \geq \frac{1}{1-a}\left[t_{k+1}^2 - a t_{k+1} + \left(\frac{1}{4} - b\right)k\right]\sigma_{k+1}(z) \\ &\qquad + \frac{\ell}{2(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left\|x^{k+1} - x^k\right\|^2 \\ &\qquad + \frac{\ell}{2(1-a)} \sum_{p=1}^{k} \left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right] \left\|x^p - x^{p-1}\right\|^2 \\ &\qquad + \frac{\ell}{2}\rho_k(z) + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2 \left\|x^{k+1} - y^{k+1}\right\|^2. \end{aligned}$$

Lemma 5.5 (i) with $k = 0$ and $y^1 = x^0$ and (5.11) with $(v^1, v^2, v^3) = (x^0, x^1, z)$ lead

to

$$\sigma_1(z) \leq -\frac{\ell}{2}\left[\left\|x^1 - z\right\|^2 - \left\|x^0 - z\right\|^2\right] - \frac{\ell - L}{2}\left\|x^1 - y^1\right\|^2.$$

From the above two inequalities and the fact that $\ell \geq L$, we can derive the desired inequality. □

Let us define the linear function $P\colon \mathbf{R} \to \mathbf{R}$ and quadratic ones $Q_1\colon \mathbf{R} \to \mathbf{R}$, $Q_2\colon \mathbf{R} \to \mathbf{R}$, and $Q_3\colon \mathbf{R} \to \mathbf{R}$ by

$$\begin{aligned} P(\alpha) &\coloneqq \frac{a^2(\alpha - 1)}{2}, \\ Q_1(\alpha) &\coloneqq \frac{1-a}{4}\alpha^2 + \left[1 - \frac{a}{2} + \frac{1-4b}{4(1-a)}\right]\alpha + 1, \\ Q_2(\alpha) &\coloneqq \frac{a(1-a)}{4}\alpha^2 + \left[\frac{a}{2} + \frac{1-4b}{4(1-a)}\right]\alpha, \\ Q_3(\alpha) &\coloneqq \left(\frac{1-a}{2}\alpha + 1\right)^2. \end{aligned} \tag{5.16}$$

The following lemma provides the key relation to evaluate the convergence rate of Algorithm 5.1.

Lemma 5.8
Under Assumption 4.1, Algorithm 5.1 generates a sequence $\{x^k\}$ *such that*

$$\begin{aligned} \frac{\ell R}{2} \geq Q_1(k)u_0(x^{k+1}) + \frac{\ell}{2}Q_2(k)\left\|x^{k+1} - x^k\right\|^2 + \frac{\ell}{2}\sum_{p=1}^{k} P(p)\left\|x^p - x^{p-1}\right\|^2 \\ + \frac{\ell - L}{2}\sum_{p=1}^{k} Q_3(p)\left\|x^{p+1} - y^{p+1}\right\|^2 \end{aligned}$$

for all $k \geq 0$, *where* $R \geq 0$ *and* $P, Q_1, Q_2, Q_3\colon \mathbf{R} \to \mathbf{R}$ *are given in* (4.8) *and* (5.16), *respectively, and* u_0 *is the gap function defined by* (3.1).

Proof. Let $k \geq 0$. With similar arguments used in the proof of Theorem 4.5 (see [**Tanabe2022a**]), we get

$$\sup_{F^* \in F(X^* \cap \mathbf{lev}_F(F(x^0)))} \inf_{z \in F^{-1}(\{F^*\})} \sigma_{k+1}(z) = u_0(x^{k+1}).$$

Since $\rho_k(z) \geq 0$, Lemma 5.7 and the above equality lead to

$$\begin{aligned}\frac{\ell R}{2} \geq{}& \frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4}-b\right)k\right] u_0(x^{k+1}) \\ &+ \frac{\ell}{2(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4}-b\right)k\right]\left\|x^{k+1}-x^k\right\|^2 \\ &+ \frac{\ell}{2(1-a)}\sum_{p=1}^{k}\left[a^2(t_p-1) + \left(\frac{1}{4}-b\right)\frac{p-t_p+a(t_p-1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|^2 \\ &+ \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2\left\|x^{p+1}-y^{p+1}\right\|^2.\end{aligned}$$

We now show that the coefficients of the four terms on the right-hand side can be bounded from below by the polynomials given in (5.16). First, by using the relation

$$t_{k+1} \geq \frac{1-a}{2}k + 1 \tag{5.17}$$

obtained from Lemma 5.3 (i) and $a \in [0,1)$, we have

$$\begin{aligned}\frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4}-b\right)k\right] &= \frac{1}{1-a}\left[t_{k+1}(t_{k+1}-a) + \left(\frac{1}{4}-b\right)k\right] \\ &\geq \frac{1}{1-a}\left[\left(\frac{1-a}{2}k+1\right)\left(\frac{1-a}{2}k+1-a\right) + \left(\frac{1}{4}-b\right)k\right] = Q_1(k).\end{aligned}$$

Again, (5.17) gives

$$\begin{aligned}\frac{1}{1-a}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4}-b\right)k\right] &= \frac{a}{1-a}t_{k+1}(t_{k+1}-1) + \frac{1-4b}{4(1-a)}k \\ &\geq \frac{a}{1-a}\left(\frac{1-a}{2}k+1\right)\left(\frac{1-a}{2}k\right) + \frac{1-4b}{4(1-a)}k = Q_2(k).\end{aligned}$$

Moreover, since $t_p \leq p$ (cf. Lemma 5.3 (ii)), $t_k \geq 1$ (cf. Lemma 5.3 (i)), and $b \in (a^2/4, 1/4]$, we obtain

$$\frac{1}{1-a}\left[a^2(t_p-1) + \left(\frac{1}{4}-b\right)\frac{p-t_p+a(t_p-1)}{t_p}\right] \geq \frac{a^2}{1-a}(t_p-1) \geq P(p).$$

It is also clear from (5.17) that

$$t_{p+1}^2 \geq Q_3(p).$$

Thus, combining the above five inequalities, we get the desired inequality. □

Then, we can finally prove the main theorem.

Theorem 5.4 (ii). It is clear from Lemma 5.8 and $Q_1(k) = O(k^2)$ as $k \to \infty$. □

Remark 5.2
Lemma 5.8 also implies the following other claims than Theorem 5.4 (ii):

- $O(1/k^2)$ *convergence rate of* $\left\{\left\|x^{k+1} - x^k\right\|^2\right\}$ *when* $a > 0$*;*
- *the absolute convergence of* $\left\{k\left\|x^{k+1} - x^k\right\|^2\right\}$ *when* $a > 0$*;*
- *the absolute convergence of* $\left\{k^2\left\|x^k - y^k\right\|^2\right\}$ *when* $\ell > L$*.*

*Note that the second one generalize [***Chambolle2015***] for single-objective problems.*

5.4 Convergence of the iterates

While the last section shows that Algorithm 5.1 has an $O(1/k^2)$ convergence rate like Algorithm 5.1, this section proves the following theorem:

Theorem 5.9
Let $\{x^k\}$ *be generated by Algorithm 5.1 with* $a > 0$*. Then, under Assumption 4.1, the following two properties hold:*

(i) $\{x^k\}$ *is bounded, and it has an accumulation point;*

(ii) $\{x^k\}$ *converges to a weak Pareto optimum for (1.1).*

The latter claim is also significant in application. For example, finite-time manifold (active set) identification, which detects the low-dimensional manifold where the optimal solution belongs, essentially requires only the convergence of the generated sequence to a unique point rather than the strong convexity of the objective functions [**Sun2019**].

Again, we will prove Theorem 5.9 after showing some lemmas. First, we mention the following result, obvious from Assumption 4.1 and Theorem 5.4 (i).

Lemma 5.10
Let $\{x^k\}$ be generated by Algorithm 5.1 and let X^* be the set of weakly Pareto optimal points. Then, for any $k \geq 0$, there exists $z \in X^* \cap \mathbf{lev}_F(F(x^0))$ (see (2.2) for the definition of $\mathbf{lev}_F$) such that

$$\sigma_k(z) \geq 0 \quad \text{and} \quad \left\|z - x^0\right\|^2 \leq R,$$

where $R \geq 0$ is given by (4.8).

The following lemma also contributes strongly to the proof of the main theorem.

Lemma 5.11
Let $\{\gamma_q\}$ be defined by line 6 in Algorithm 5.1. Then, we have

$$\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q \leq 2(s-1) \quad \text{for all } s, r \geq 1.$$

Proof. By using Lemma 5.3 (iv), we see that

$$\prod_{q=s}^{p} \gamma_q \leq \prod_{q=s}^{p} \frac{q-1}{q+1/2}.$$

Let Γ and B denote the gamma and beta functions defined by

$$\Gamma(\alpha) := \int_0^\infty \tau^{\alpha-1} \exp(-\tau)\, \mathrm{d}\tau \quad \text{and} \quad \mathrm{B}(\alpha, \beta) := \int_0^1 \tau^{\alpha-1}(1-\tau)^{\beta-1}\, \mathrm{d}\tau, \tag{5.18}$$

respectively. Applying the well-known properties:

$$\Gamma(\alpha) = (\alpha-1)!, \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \text{and} \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{5.19}$$

we get

$$\prod_{q=s}^{p} \gamma_q \leq \frac{\Gamma(p)/\Gamma(s-1)}{\Gamma(p+3/2)/\Gamma(s+1/2)} = \frac{B(p, 3/2)}{B(s-1, 3/2)}.$$

This implies

$$\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q \leq \sum_{p=1}^{r} B(p, 3/2)/B(s-1, 3/2).$$

Then, it follows from the definition (5.18) of B that

$$\begin{aligned}\sum_{p=s}^{r}\prod_{q=s}^{p}\gamma_q &\le \sum_{p=s}^{r}\int_0^1 \tau^{p-1}(1-\tau)^{1/2}\,\mathrm{d}\tau/B(s-1,3/2)\\ &= \int_0^1 \sum_{p=s}^{r}\tau^{p-1}(1-\tau)^{1/2}\,\mathrm{d}\tau/B(s-1,3/2)\\ &= \int_0^1 \frac{\tau^{s-1}-\tau^{r}}{1-\tau}(1-\tau)^{1/2}\,\mathrm{d}\tau/B(s-1,3/2)\\ &= \frac{B(s,1/2)-B(r+1,1/2)}{B(s-1,3/2)} \le \frac{B(s,1/2)}{B(s-1,3/2)}.\end{aligned}$$

Using again (5.19), we conclude that

$$\sum_{p=s}^{r}\prod_{q=s}^{p}\gamma_q \le \frac{\Gamma(s)\Gamma(1/2)/\Gamma(s+1/2)}{\Gamma(s-1)\Gamma(3/2)/\Gamma(s+1/2)} = 2(s-1). \qquad \square$$

Now, we introduce two functions $\omega_k\colon \mathbf{R}^n \to \mathbf{R}$ and $\nu_k\colon \mathbf{R}^n \to \mathbf{R}$ for any $k \ge 1$, which will help our analysis, by

$$\omega_k(z) \coloneqq \max\left(0, \left\|x^k - z\right\|^2 - \left\|x^{k-1} - z\right\|^2\right), \tag{5.20}$$

$$\nu_k(z) \coloneqq \left\|x^k - z\right\|^2 - \sum_{s=1}^{k}\omega_s(z). \tag{5.21}$$

The lemma below describes the properties of ω_k and ν_k.

Lemma 5.12
Let $\{x^k\}$ be generated by Algorithm 5.1 and recall that $\mathbf{lev}_F, \omega_k$, and ν_k are defined by (2.2), (5.20) and (5.21), respectively. Moreover, suppose that Assumption 4.1 holds and that $z \in X^* \cap \mathbf{lev}_F(F(x^0))$ satisfies the statement of Lemma 5.10 for some $k \ge 1$. Then, it follows for all $r = 1, \dots, k$ that

(i) $\displaystyle\sum_{s=1}^{r}\omega_s(z) \le \sum_{s=1}^{r}(6s-5)\left\|x^s - x^{s-1}\right\|^2;$

(ii) $\nu_{r+1}(z) \le \nu_r(z).$

Proof. (i) : Let $k \ge p \ge 1$. From the definition of y^{p+1} given in line 7 of Algo-

rithm 5.1, we have

$$\begin{aligned}
&\left\|x^{p+1}-z\right\|^2-\left\|x^p-z\right\|^2\\
&=-\left\|x^{p+1}-x^p\right\|^2+2\left\langle x^{p+1}-y^{p+1},x^{p+1}-z\right\rangle+2\gamma_p\left\langle x^p-x^{p-1},x^{p+1}-z\right\rangle\\
&=-\left\|x^{p+1}-x^p\right\|^2+2\left\langle x^{p+1}-y^{p+1},y^{p+1}-z\right\rangle+2\left\|x^{p+1}-y^{p+1}\right\|^2\\
&\quad+2\gamma_p\left\langle x^p-x^{p-1},x^{p+1}-z\right\rangle.
\end{aligned}$$

On the other hand, Lemma 5.5 (i) gives

$$2\left\langle x^{p+1}-y^{p+1},y^{p+1}-z\right\rangle\le-\frac{2}{\ell}\sigma_{p+1}(z)-\frac{2\ell-L}{\ell}\left\|x^{p+1}-y^{p+1}\right\|^2.$$

Moreover, Lemma 5.6 with $(k_1,k_2)=(p+1,k+1)$ implies

$$\begin{aligned}
&-\frac{2}{\ell}\sigma_{p+1}(z)\\
&\quad\le-\frac{2}{\ell}\sigma_{k+1}(z)-\left\|x^{k+1}-x^k\right\|^2+\left\|x^{p+1}-x^p\right\|^2-\sum_{r=p+1}^{k}\frac{1}{t_r}\left\|x^r-x^{r-1}\right\|^2\\
&\quad\le\left\|x^{p+1}-x^p\right\|^2,
\end{aligned}$$

where the second inequality comes from the assumption on z. Combining the above three inequalities, we get

$$\begin{aligned}
&\left\|x^{p+1}-z\right\|^2-\left\|x^p-z\right\|^2\le\frac{L}{\ell}\left\|x^{p+1}-y^{p+1}\right\|^2+2\gamma_p\left\langle x^p-x^{p-1},x^{p+1}-z\right\rangle\\
&=\frac{L}{\ell}\left\|x^{p+1}-y^{p+1}\right\|^2+\gamma_p\Big(\left\|x^p-z\right\|^2-\left\|x^{p-1}-z\right\|^2+\left\|x^p-x^{p-1}\right\|^2\\
&\qquad\qquad+2\left\langle x^p-x^{p-1},x^{p+1}-x^p\right\rangle\Big).
\end{aligned}$$

Using the relation $\|x^{p+1}-y^{p+1}\|^2+2\gamma_p\langle x^p-x^{p-1},x^{p+1}-x^p\rangle=\|x^{p+1}-x^p\|^2+\gamma_p^2\|x^p-x^{p-1}\|^2$, which holds from the definition of y^k, we have

$$\begin{aligned}
\left\|x^{p+1}-z\right\|^2-\left\|x^p-z\right\|^2\le&-\frac{\ell-L}{\ell}\left\|x^{p+1}-y^{p+1}\right\|^2+\left\|x^{p+1}-x^p\right\|^2\\
&+\gamma_p\left(\left\|x^p-z\right\|^2-\left\|x^{p-1}-z\right\|^2\right)+(\gamma_p+\gamma_p^2)\left\|x^p-x^{p-1}\right\|^2.
\end{aligned}$$

Since $0 \le \gamma_p \le 1$ from Lemma 5.3 (iv) and $\ell \ge L$, we obtain

$$\begin{aligned}\left\|x^{p+1}-z\right\|^2 - \left\|x^p - z\right\|^2 &\\ \le \gamma_p \left(\left\|x^p - z\right\|^2 - \left\|x^{p-1}-z\right\|^2 + 2\left\|x^p - x^{p-1}\right\|^2\right) + \left\|x^{p+1}-x^p\right\|^2 &\\ \le \gamma_p \left(\omega_p(z) + 2\left\|x^p - x^{p-1}\right\|^2\right) + \left\|x^{p+1}-x^p\right\|^2,&\end{aligned}$$

where the second inequality follows from the definition (5.20) of ω_p. Since the right-hand side is nonnegative, (5.20) again gives

$$\omega_{p+1}(z) \le \gamma_p \left(\omega_p(z) + 2\left\|x^p - x^{p-1}\right\|^2\right) + \left\|x^{p+1}-x^p\right\|^2.$$

Let $s \le k$. Applying the above inequality recursively and using $\gamma_1 = 0$, we get

$$\begin{aligned}\omega_s(z) &\le 3\sum_{p=2}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p - x^{p-1}\right\|^2 + 2\prod_{q=1}^{s}\gamma_q\left\|x^1 - x^0\right\|^2 + \left\|x^s - x^{s-1}\right\|^2 \\ &\le 3\sum_{p=2}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p - x^{p-1}\right\|^2 + \left\|x^s - x^{s-1}\right\|^2.\end{aligned}$$

Adding up the above inequality from $s = 1$ to $s = r \le k$, we have

$$\begin{aligned}\sum_{s=1}^{r}\omega_s(z) &\le 3\sum_{s=1}^{r}\sum_{p=1}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p - x^{p-1}\right\|^2 + \sum_{s=1}^{r}\left\|x^s - x^{s-1}\right\|^2 \\ &= 3\sum_{p=1}^{r}\sum_{s=p}^{r}\prod_{q=p}^{s}\gamma_q\left\|x^p - x^{p-1}\right\|^2 + \sum_{s=1}^{r}\left\|x^s - x^{s-1}\right\|^2 \\ &= \sum_{s=1}^{r}\left(3\sum_{p=s}^{r}\prod_{q=s}^{p}\gamma_q + 1\right)\left\|x^s - x^{s-1}\right\|^2,\end{aligned}$$

where the first equality follows from (5.12). Thus, Lemma 5.11 implies

$$\sum_{s=1}^{r}\omega_s(z) \le \sum_{s=1}^{r}(6s-5)\left\|x^s - x^{s-1}\right\|^2.$$

(ii) : Eq. (5.21) yields

$$\begin{aligned}\nu_{r+1}(z) &= \left\|x^{r+1}-z\right\|^2 - \omega_{r+1}(z) - \sum_{s=1}^{r}\omega_s(z)\\ &= \left\|x^{r+1}-z\right\|^2 - \max\left(0, \left\|x^{r+1}-z\right\|^2 - \|x^r-z\|^2\right) - \sum_{s=1}^{r}\omega_s(z)\\ &\le \left\|x^{r+1}-z\right\|^2 - \left(\left\|x^{r+1}-z\right\|^2 - \|x^r-z\|^2\right) - \sum_{s=1}^{r}\omega_s(z)\\ &= \|x^r-z\|^2 - \sum_{s=1}^{r}\omega_s(z) = \nu_r(z),\end{aligned}$$

where the second and third equalities come from the definitions (5.20) and (5.21) of ω_{r+1} and ν_r, respectively. □

Let us now prove the first part of the main theorem.

Theorem 5.9 (i). Let $k \ge 1$ and suppose that $z \in X^* \cap \mathbf{lev}_F(F(x^0))$ satisfies the statement of Lemma 5.10, where X^* is the set of weakly Pareto optimal solutions and $\mathbf{lev}_F$ is given by (2.2). Then, Lemma 5.12 (ii) gives

$$\begin{aligned}\nu_k(z) \le \nu_1(z) &= \left\|x^1-z\right\|^2 - \omega_1(z)\\ &= \left\|x^1-z\right\|^2 - \max\left(0, \left\|x^1-z\right\|^2 - \left\|x^0-z\right\|^2\right)\\ &\le \left\|x^1-z\right\|^2 - \left(\left\|x^1-z\right\|^2 - \left\|x^0-z\right\|^2\right) = \left\|x^0-z\right\|^2,\end{aligned}$$

where the second equality follows from the definition (5.20) of ω_1. Considering the definition (5.21) of ν_k, we obtain

$$\left\|x^k-z\right\|^2 \le \left\|x^0-z\right\|^2 + \sum_{s=1}^{k}\omega_s(z).$$

Taking the square root of both sides and using (5.20), we get

$$\left\|x^k-z\right\| \le \sqrt{\|x^0-z\|^2 + \sum_{s=1}^{k}(6s-5)\|x^s-x^{s-1}\|^2}.$$

Applying the reverse triangle inequality $\left\|x^k-x^0\right\| - \|x^0-z\| \le \left\|x^k-z\right\|$ to the

left-hand side leads to

$$\begin{aligned}\left\|x^k - x^0\right\| &\leq \left\|x^0 - z\right\| + \sqrt{\|x^0 - z\|^2 + \sum_{s=1}^{k}(6s-5)\|x^s - x^{s-1}\|^2} \\ &\leq \sqrt{R} + \sqrt{R + \sum_{s=1}^{k}(6s-5)\|x^s - x^{s-1}\|^2},\end{aligned}$$

where the second inequality comes from the assumption on z. Moreover, since $a > 0$, the right-hand side is bounded from above according to Lemma 5.8. This implies that $\left\{x^k\right\}$ is bounded, and so it has accumulation points. □

Before proving Theorem 5.9 (ii), we show the following lemma.

Lemma 5.13

Let $\left\{x^k\right\}$ be generated by Algorithm 5.1 with $a > 0$ and suppose that Assumption 4.1 holds. Then, if $\bar{z}$ is an accumulation point of $\left\{x^k\right\}$, then $\left\{\left\|x^k - \bar{z}\right\|\right\}$ is convergent.

Proof. Assume that $\left\{x^{k_j}\right\} \subseteq \left\{x^k\right\}$ converges to $\bar{z}$. Then, we have $\sigma_{k_j}(\bar{z}) \to 0$ by the definition (5.10) of σ_{k_j}. Therefore, we can regard $\bar{z}$ to satisfy the statement of Lemma 5.10 at $k = \infty$, and thus the inequalities of Lemma 5.12 hold for any $r \geq 1$ and $\bar{z}$. This means $\{\nu_k(\bar{z})\}$ is non-increasing and bounded, i.e., convergent. Hence $\left\{\left\|x^k - \bar{z}\right\|\right\}$ is convergent. □

Finally, we finish the proof of the main theorem.

Theorem 5.9 (ii). Suppose that $\left\{x^{k_j^1}\right\}$ and $\left\{x^{k_j^2}\right\}$ converges to $\bar{z}^1$ and $\bar{z}^2$, respectively. From Lemma 5.13, we see that

$$\lim_{j\to\infty}\left(\left\|x^{k_j^2} - \bar{z}^1\right\|^2 - \left\|x^{k_j^2} - \bar{z}^2\right\|^2\right) = \lim_{j\to\infty}\left(\left\|x^{k_j^1} - \bar{z}^1\right\|^2 - \left\|x^{k_j^1} - \bar{z}^2\right\|^2\right).$$

This yields that $\|\bar{z}^1 - \bar{z}^2\|^2 = -\|\bar{z}^1 - \bar{z}^2\|^2$, and so $\|\bar{z}^1 - \bar{z}^2\|^2 = 0$, i.e., $\left\{x^k\right\}$ is convergent. Let $x^k \to x^*$. Since $\left\|x^{k+1} - x^k\right\|^2 \to 0$, $\left\{y^k\right\}$ is also convergent to x^*. Therefore, Proposition 5.2 shows that x^* is weakly Pareto optimal for (1.1). □

5.5 Numerical experiments

This section compares the performance between Algorithm 5.1 with various a and b and Algorithm 5.1 $(a = 0, b = 1/4)$ through numerical experiments. We run all

experiments in Python 3.9.9 on a machine with 2.3 GHz Intel Core i7 CPU and 32 GB memory. For each example, we test 15 different hyperparameters combining $a = 0, 1/6, 1/4, 1/2, 3/4$ and $b = a^2/4, (a^2+1)/8, 1/4$, i.e.,

$$(a,b) = \left\{ \begin{array}{c} (0,0), (0,1/8), (0,1/4), \\ (1/6,1/144), (1/6,37/288), (1/6,1/4), \\ (1/4,1/64), (1/4,17/128), (1/4,1/4), \\ (1/2,1/16), (1/2,5/32), (1/2,1/4), \\ (3/4,9/64), (3/4,25/128), (3/4,1/4) \end{array} \right\},$$

and we set $\varepsilon = 10^{-5}$ for the stopping criteria.

5.5.1 Artificial test problems (bi-objective and tri-objective)

First, we solve the multi-objective test problems in the form (1.1) used in [**Tanabe2022a**], modifications from [**Jin2001**, **Fliege2009**], whose objective functions are defined by

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = g_2(x) = 0, \tag{JOS1}$$

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = \frac{1}{n}\|x\|_1, g_2(x) = \frac{1}{2n}\|x-1\|_1, \tag{JOS1-L1}$$

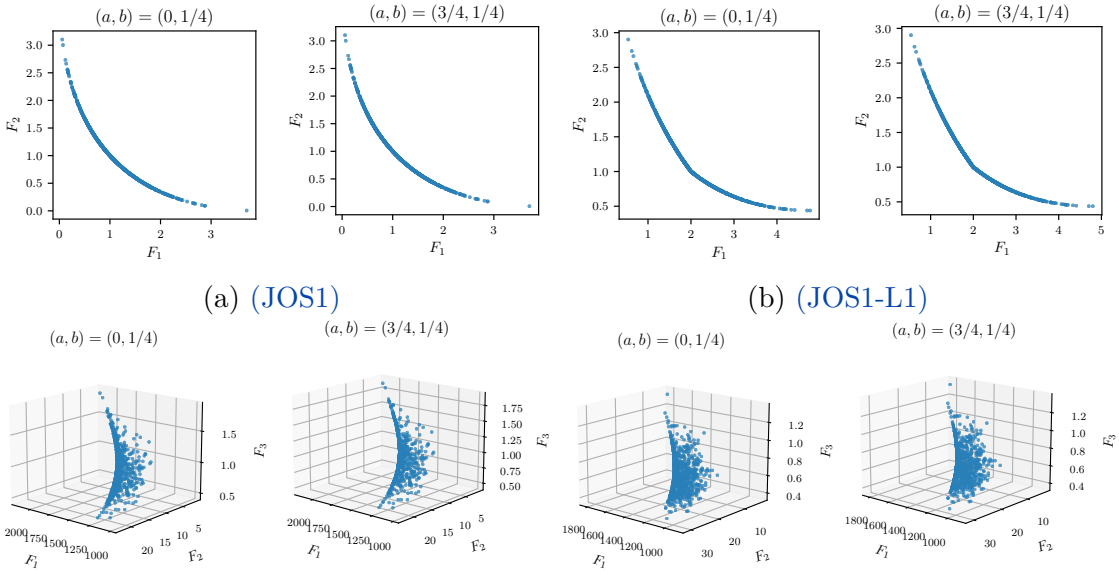
$$\left\{ \begin{array}{l} f_1(x) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \dfrac{1}{n(n+1)}\displaystyle\sum_{i=1}^{n} i(n-i+1)\exp(-x_i), g_1(x) = g_2(x) = g_3(x) = 0, \end{array} \right. \tag{FDS}$$

$$\left\{ \begin{array}{l} f_1(x) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \dfrac{1}{n(n+1)}\displaystyle\sum_{i=1}^{n} i(n-i+1)\exp(-x_i), g_1(x) = g_2(x) = g_3(x) = \delta_{\mathbf{R}^n_+}(x), \end{array} \right. \tag{FDS-CON}$$

where $x \in \mathbf{R}^n, n = 50$ and $\delta_{\mathbf{R}^n_+}$ is an indicator function (1.7) of the nonnegative orthant. We choose 1000 initial points, commonly for all pairs (a,b), and randomly with a uniform distribution between $\underline{c}$ and $\overline{c}$, where $\underline{c} = (-2,\ldots,-2)^\top$

and $\overline{c} = (4, \ldots, 4)^\top$ for (JOS1) and (JOS1-L1), $\underline{c} = (-2, \ldots, -2)^\top$ and $\overline{c} = (2, \ldots, 2)^\top$ for (FDS), and $\underline{c} = (0, \ldots, 0)^\top$ and $\overline{c} = (2, \ldots, 2)^\top$ for (FDS-CON). Moreover, we use backtracking for updating ℓ, with 1 as the initial value of ℓ and 2 as the constant multiplied into ℓ at each iteration (cf. [**Tanabe2022a**]). Furthermore, at each iteration, we transform the subproblem (5.1) into their dual as suggested in [**Tanabe2022a**] and solve them with the trust-region interior point method [**Byrd1999**] using the scientific library SciPy.

Figure 5.1 and Table 5.1 present the experimental results. Figure 5.1 plots the solutions only for the cases $(a, b) = (0, 1/4), (3/4, 1/4)$, but other combinations also yield similar plots, including a wide range of Pareto solutions. Table 5.1 shows that the new momentum factors are fast enough to compete with the existing ones ($(a, b) = (0, 1/4)$ or $b = a^2/4$) and better than them in some cases.



(a) (JOS1) (b) (JOS1-L1)

(c) (FDS) (d) (FDS-CON)

Figure 5.1: Pareto solutions obtained with some (a, b)

5.5.2 Image deblurring (single-objective)

Since our proposed momentum factor is also new in the single-objective context, we also tackle deblurring the cameraman test image via a single-objective ℓ_2-ℓ_1 minimization, inspired by [**Beck2009**]. In detail, as shown in Figure 5.2, to a 256×256 cameraman test image with each pixel scaled to $[0, 1]$, we generate an observed

Table 5.1: Average computational costs to solve the multi-objective examples

(a) (JOS1)

a	b	Time [s]	Iterations
0	0	6.442	97.0
0	1/8	5.158	81.217
0	1/4	4.207	65.0
1/6	1/144	4.244	67.0
1/6	37/288	5.182	82.0
1/6	1/4	4.268	66.0
1/4	1/64	6.224	99.0
1/4	17/128	7.239	113.566
1/4	1/4	3.205	51.0
1/2	1/16	4.51	72.0
1/2	5/32	4.562	71.0
1/2	1/4	4.466	70.0
3/4	9/64	4.323	67.998
3/4	25/128	3.104	49.0
3/4	1/4	3.741	47.0

(b) (JOS1-L1)

a	b	Time [s]	Iterations
0	0	10.733	157.512
0	1/8	11.054	161.065
0	1/4	11.122	161.734
1/6	1/144	9.85	141.731
1/6	37/288	9.994	144.863
1/6	1/4	10.399	150.592
1/4	1/64	9.271	135.804
1/4	17/128	9.463	137.108
1/4	1/4	9.662	139.848
1/2	1/16	7.439	109.082
1/2	5/32	7.642	110.204
1/2	1/4	7.723	111.599
3/4	9/64	5.253	77.366
3/4	25/128	5.39	79.425
3/4	1/4	5.678	82.37

(c) (FDS)

a	b	Time [s]	Iterations
0	0	29.24	204.438
0	1/8	29.797	210.595
0	1/4	30.565	214.934
1/6	1/144	24.964	174.393
1/6	37/288	25.375	177.944
1/6	1/4	26.065	182.398
1/4	1/64	22.94	159.737
1/4	17/128	23.311	162.629
1/4	1/4	23.976	166.918
1/2	1/16	17.909	122.653
1/2	5/32	18.14	123.96
1/2	1/4	18.221	125.697
3/4	9/64	13.584	94.176
3/4	25/128	13.674	94.705
3/4	1/4	13.795	94.868

(d) (FDS-CON)

a	b	Time [s]	Iterations
0	0	37.345	259.508
0	1/8	37.439	261.522
0	1/4	37.94	263.911
1/6	1/144	32.463	227.063
1/6	37/288	38.265	229.736
1/6	1/4	45.661	231.958
1/4	1/64	41.434	209.35
1/4	17/128	33.664	211.69
1/4	1/4	30.772	213.811
1/2	1/16	22.92	158.448
1/2	5/32	23.1	159.685
1/2	1/4	23.539	162.226
3/4	9/64	17.092	118.616
3/4	25/128	17.123	118.063
3/4	1/4	17.115	118.844

image by applying a Gaussian blur of size 9×9 and standard deviation 4 and adding a zero-mean white Gaussian noise with standard deviation 10^{-3}.

(a) Original (b) Blurred and noisy

Figure 5.2: Deblurring of the cameraman

Letting θ, B, and W be the observed image, the blur matrix, and the inverse of the Haar wavelet transform, respectively, consider the single-objective problem (1.1) with $m = 1$ and

$$f_1(x) \coloneqq \|BWx - \theta\|^2 \quad \text{and} \quad g_1(x) = \lambda \|x\|_1, \tag{CAM-DEBLUR}$$

where $\lambda \coloneqq 2 \times 10^{-5}$ is a regularization parameter. Unlike in the previous subsection, we can compute ∇f's Lipschitz constant by calculating $(BW)^\top (BW)$'s eigenvalues using the two-dimensional cosine transform [**Hansen2006**], so we use it constantly as ℓ. Moreover, we use the observed image's Wavelet transform as the initial point.

Figure 5.3 shows the reconstructed image from the obtained solution. Images produced by all hyperparameters are similar, so we present only $(a, b) = (0, 1/4)$ and $(1/2, 1/4)$. Moreover, we summarize the numerical performance in Table 5.2 and Figure 5.4. Like the last subsection, this example also suggests that our new momentum factors may occasionally improve the algorithm's performance even for single-objective problems.

5.6 Conclusions

We have generalized the momentum factor of the multi-objective accelerated proximal gradient algorithm [**Tanabe2022a**] in a form that is even new in the single-

(a) $(a, b) = (0, 1/4)$ (b) $(a, b) = (1/2, 1/4)$

Figure 5.3: Deblurred image

Table 5.2: Computational costs for the image deblurring

a	b	Total time [s]	Iteration counts
0	0	75.227	558
0	1/8	75.176	558
0	1/4	75.388	558
1/6	1/144	66.499	460
1/6	37/288	66.866	462
1/6	1/4	66.685	462
1/4	1/64	61.791	421
1/4	17/128	61.622	421
1/4	1/4	35.69	421
1/2	1/16	26.828	306
1/2	5/32	26.274	304
1/2	1/4	25.535	303
3/4	9/64	32.54	369
3/4	25/128	30.473	364
3/4	1/4	27.713	360

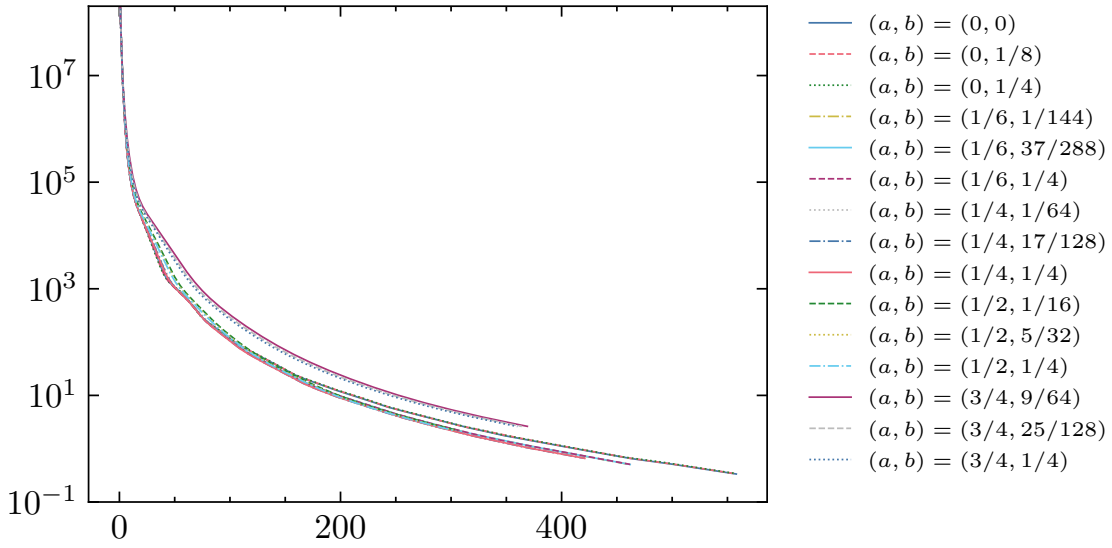


Figure 5.4: Values of $u_0(x^k) = F_1(x) - F_1(x^*)$, where x^* is the optimal solution estimated from the original image

objective context and includes the known FISTA momentum factors [**Beck2009**, **Chambolle2015**]. Furthermore, with the proposed momentum factor, we proved under reasonable assumptions that the algorithm has an $O(1/k^2)$ convergence rate and that the iterates converge to Pareto solutions. Moreover, the numerical results reinforced these theoretical properties and suggested the potential for our new momentum factor to improve the performance.

Our proposed method has the potential to achieve finite-time manifold (active set) identification [**Sun2019**] without the assumption of the strong convexity (or its generalizations such as PL conditions or error bounds [**Karimi2016**]). Moreover, we took a single update rule of t_k for all iterations in this work, but the adaptive change of the strategy in each iteration is conceivable. It might also be interesting to estimate the Lipschitz constant simultaneously with that change, like in [**Scheinberg2014**]. In addition, an extension to the inexact scheme like [**Villa2013**] would be significant. Those are issues to be addressed in the future.

Chapter 6

Conclusions

This thesis has proposed new merit functions, the proximal gradient method, and the accelerated proximal gradient method for non-smooth multi-objective optimization problems. We summarize the results obtained here as follows:

(i) In Chapter 3, we have proposed three merit functions for non-smooth multi-objective optimization: (i) the gap function for continuous multi-objective optimization; (ii) the regularized gap function for convex multi-objective optimization; (iii) the regularized and partially linearized gap function for composite multi-objective optimization. First, we have shown that they actually satisy the propeties as merit functions and proved the lower semi-continuity of (i) and the locally Lipschitz continuity of (ii) and (iii). We have also confirmed the differentiability of (ii) and (iii) under reasonable assumptions and that the stationary points of (ii) and (iii) solve the original multi-objective problem under strict convexity. Secondly, we have derived inequalities among different merit functions under certain conditions. We thirdly have demonstrated that the level-boundedness of the objective functions implies the level-boundedness of the associated merit functions, and we finally proved that the objective functions' strong convexity provides the error bound property of the merit functions.

(ii) In Chapter 4, we have developed the proximal gradient method for composite multi-objective optimization. We have shown that every accumulation point of the generated sequence, if it exists, is Pareto stationary. Moreover, we presented global convergence rates for the proposed algorithm, matching what we know in scalar optimization for non-convex $O(\sqrt{1/k})$, convex $O(1/k)$, strongly

convex $O(r^k)$ for some $r \in (0, 1)$. We also have extended the so-called Polyak-Łojasieqicz (PL) inequality for multi-objective optimziation and established the linear convergence rate for multi-objecvive problems that satisfy such inequalities. Furthermore, we have converted the subproblems to well-known convex optimization problems for robust multi-objective problem. Finally, we have reported some numerical results.

(iii) In Chapter 5, we have proposed the accelerated proximal gradient method for convex composite multi-objective optimization. We have proved the proposed methods' $O(1/k^2)$ convergence rate, together with the global convergence property. This method includes some hyperparameters, which is new even for single-objective cases. We finally have reported some numerical results, showing that some of these choices give better results than the classical algorithms.

We believe that these contributions have had some impact on non-smooth and composite multi-objective optimization. However, there are still many open problems. We conclude this thesis by describing future works related to our results.

(i) We can consider our proposed merit function's natural extension to infinite-dimensional vector optimization. We can also regard other famous merit functions' generalization to multi-objective or vector problems, such as the implicit Lagrangian and the squared Fischer-Burmeister function. Moreover, it would be interesting to develop new multi-objective algorithm using such merit functions.

(ii) Extending the many variants of the proximal gradient method in single-objective optimization to multi-objective optimization problems is a challenge that needs addressing. Obtaining a theoretically sound extension will not be straightforward for any method. However, we believe that finding practical applications of composite multi-objective optimization, such as machine learning, will significantly impact this field.

(iii) It is also crucial to extend the variants of the accelerated proximal gradient method to multi-objective optimization. Moreover, applying our acceleration techniques to large-scale problems like stochastic accelerated gradient descent would be interesting. Developing internal techniques, such as a warm start for subproblems and inexact methods, would also be necessary for applications.